Introspection and Communication: A Game-Theoretic Approach

by

Drazen Prelec

88-032

Harvard University

April, 1987

Acknowledgments: This work was supported by a Junior Fellowship from the Harvard Society of Fellows, the Milton Foundation, and the Research Division of the Harvard Graduate School of Business Administration. The author is grateful to Richard J. Herrnstein, R. Duncan Luce, Andrew McLellan, and Barry Nalebuff for valuable discussions of the ideas presented here.

ABSTRACT

Introspective reports are a means of communicating information about internal events and processes, and their success depends both on the encoding skills of the introspecting observer, and on the decoding skills of the target audience. The paper proposes a formal test by which a complete outsider to this communicative exchange can measure how much genuine information is being passed. The test requires two observer subjects to play a zero-sum game against a third, non-observing subject, who attempts to simulate the judgments of the observers, but without benefit of exposure to the stimulus. The strategic role of this third, "skeptical" subject is to disprove that the other two are indeed communicating. Importantly, the scores in the game to not depend on any objective stimulus property, but are functions of responses, only. It is shown that if everyone plays so as to maximize expected score, then: (1) the two observers will identify essentially all subjectively noticeable aspects of the stimulus (in a vocabulary of their choosing); and (2) their score will equal the amount of transmitted information (measured in bits). These results hold even if the game is played only once, with a single, undifferentiated stimulus. Like single-detection methods, the game enables one to train introspective observers through contingent payoffs, but, unlike SD, it does not oblige the experimenter to define the set of possible stimulus descriptions, nor to decide what description is "correct." The proposed method is illustrated with some experimental results, obtained with a variant game (the Information Pump) that elicits the introspective information in the form of a sequence of True/False statements.

TABLE OF CONTENTS

1 INTRODUCTION

The Problem

According to wine writer Pamela Vandyke Price, the experience of tasting wine can be recorded by a line drawing, such as the one presented below:

The figure traces the temporal unfolding of sensation, from first contact to aftertaste. Although the vertical dimension corresponds roughly to agreeable intensity of flavor, a full decoding of a particular drawing is a fairly complex matter, requiring one to attend to a variety of pictorial cues, such as smoothness, roundness, and so on. Thus, for example, the profile shown above is characteristic of a "great wine at its peak;" in contrast, a "dull wine" might look something like this (Vandyke Price, 1975, p.44; cited in Lehrer, 1983):

Vandyke Price's drawings are a type of introspective report, distinguished from others by their unusual nature. Like any form of introspective evidence, they invite a number of familiar worries and objections. To begin with the most fundamental concern, how do we know that the profiles are being constructed in a consistent manner, that, in other words, similar drawings were fathered by similar flows of sensation? Even if consistently generated, the drawings could still mislead in other ways, by leaving

out relevant aspects of wine-sensation, by introducing phenomenal aspects unrelated to wine, by creating spurious pictorial distinctions or relationships, and so on.

These, and other objections to accepting the wine-drawings as a window into the tasting experience do not arise from any doubt about the author's sincerity and effort; what troubles us is the possibility that the author herself may not know how to consistently apply the descriptive rules that she has set up. In order to do that, we note, she must have a way of discriminating correct from incorrect usage, that is, a way of choosing the one trace that most accurately corresponds to <u>that</u> sensory event. But, who is to correct her when she falls into a mistake--a misapplication of some intended rule, or an incorrect choice of drawing?

Perhaps I am attributing more weight to these impressionistic drawings then they are intended to bear. However, in puzzling over what, if anything the drawings are conveying to their audience, our attention is brought to focus on a classical problem in psychological method. In its investigation of mental activity, psychology finds it necessary to elicit, and theoretically organize descriptions of events that are normally observable by only one person: the introspecting subject. Among these, notably, are verbal acccounts of "such things as we call feelings, desires, cognitions, reasonings, decisions," (the phenomena of mental life, according to William James), as well as more regimented judgments of subjective similarity, intensity, category membership, imagery, and the like. No less that the wine-drawings, these familiar verbal productions leave room for doubt about whether there is a sufficiently stable and transparent relation between internal event and overt report to make the latter a useful source of psychological data. Such doubt is only partially eased by intersubjective

agreement, which could, after all, result from common response tendencies, unrelated to the specific mental process we are attempting to study.

What is an Introspective Report?

Our investigation into this problem takes the form of a search for an objective criterion for assessing the informational content of introspective reports, in a controlled setting. Specifically, we will be concerned with understanding the formal properties of a game-like procedure that can discriminate among more-or-less informative ways of describing private events. At the end of this section I will outline the main features of this procedure, but now I would like to clarify what I mean by an introspective report, and describe the main problem that such reports present.

In ordinary speech, the term introspective is reserved for a person's descriptions of internal processes and events--his sensations, thoughts, feelings, and the like. My intention, however, is to use of the word in a somewhat broader sense. By an introspective report I mean any form of subjective judgment that the experimenter is unable to challenge, because he does not know the conditions that make that report true, or false. Such a judgment might be overtly subjective, such as:

"I feel a tingling sensation,"

or,

"Within each personal consciousness, thought is sensibly continuous;"[1]

and then again it might refer to some property of an evaluated object, as:

"This is an unusual wine,"

or,

---

1 William James; 1890/1981, p. 221.

"Kandinsky's middle period is most interesting."

The dividing line I wish to observe follows the private/public, rather than the inner/outer distinction. An estimate of one's blood pressure, for example, would not count as introspective, because blood pressure is a public fact, notwithstanding that it is an event that occurs within one's body.

Introspective judgements, in this broad sense, play an important role in psychology, as in the social sciences generally. Here are three examples of the use of introspective evidence in current cognitive research:

1. In studies of similarity, subjects are shown objects from a set and asked to rate pairs of objects by their perceived similarity (Shepard, 1974; Tversky and Gati, 1982).
2. In experiments on categorization, subjects list features of objects, they identify which objects are most typical of a category, etc. (Rosch et al., 1976).
3. The cognitive processes engaged by imagery are investigated by asking subjects to indicate how long it takes them to imagine, rotate, magnify images of various objects (Kosslyn, 1980).

Objective procedures: The signal-detection paradigm

In the 1950's, the application of the statistical theory of signal-detection to psychophysics contributed a fundamentally new insight into the problem of instructing the introspecting observer, an insight which we will take as our starting point (Tanner, Swets, and Birdsall, 1956; Green and Swets, 1966). Figure 1 gives the basic schema for the signal-detection experiment. The experimental subject observes a stimulus, and responds by selecting an element from previously specified response

set. The "correct" stimulus characteristics are then revealed, and the subject's response scored according to some (public) rule. In this procedure, no problem of referential ambiguity can arise, because whenever the subject asks:

"How should I respond if I see (or hear, or remember, or feel...) this?"

the experimenter has ready a clear reply:

"Select the response that maximizes your expected score."

Insert Figure 1 About Here

The correspondence of internal state, α, to appropriate response, a, or the "semantics" of the response categories, as it were, is established not through any direct observation of the subject's internal state -- the "this" referred to in his query, -- but through a statistical correlation of these states and the relevant, public stimulus characteristics, denoted by s in Figure 1. Barring ties in expected score, every internal state is covered by only one correct response, which the subject can learn to identify through trial-and-error if the relationship between α and s is not obvious, or if there was some obscurity in the instructions. In this way a response becomes an implicit description of those internal states for which it is the score-maximizing choice. The scoring rule carves the set of possible experiences into a small number of mutually exclusive subsets, each of which is covered by a single response.

Although the signal detection procedure was first introduced into psychology in order to settle a rather narrow psychophysical issue -- the existence of a sensory threshold -- the problem which it solved was a quite

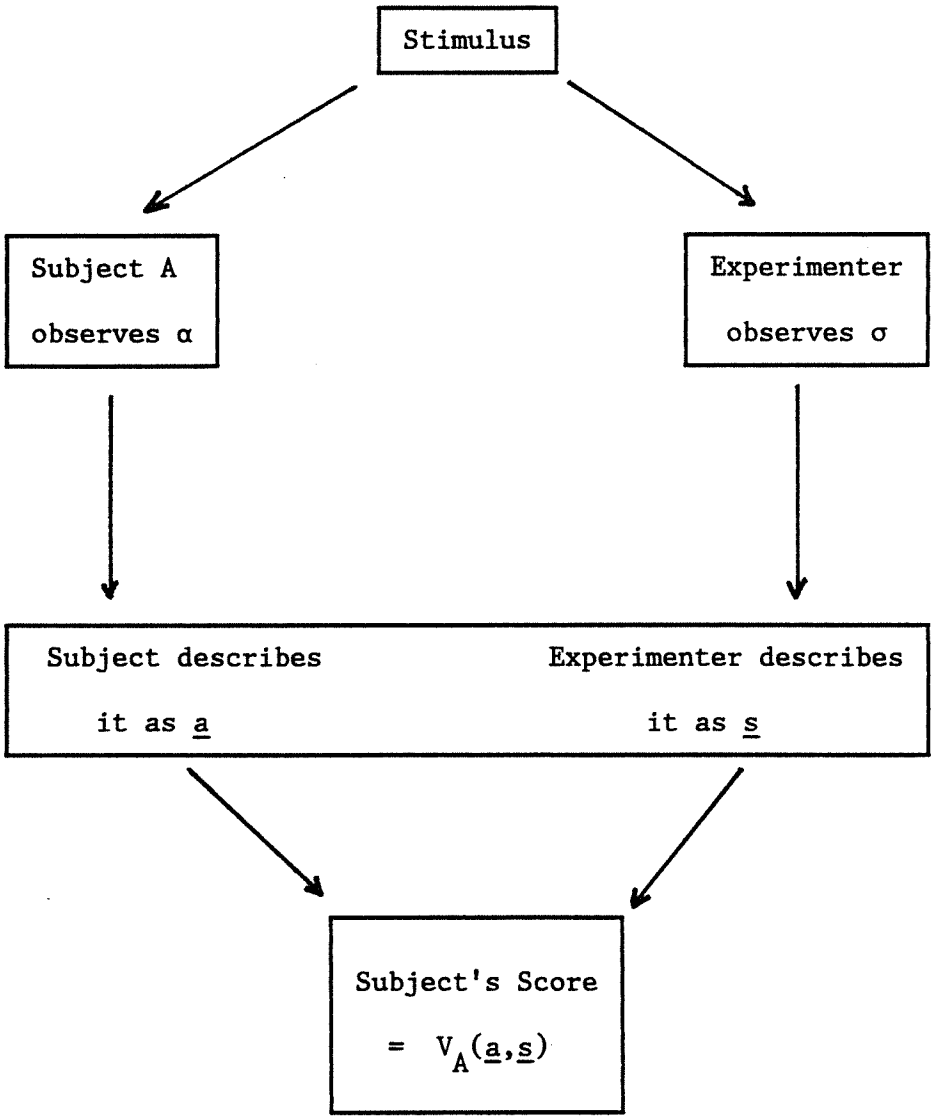


Figure 1: The signal-detection paradigm: The subject is instructed to maximize expected score.

general one (Edwards, 1961). It showed how one could, in principle, instruct a person to consistently pick out and report any phenomenal aspect that contributes to a person's knowledge of public stimulus characteristics.

The referential clarity provided by this paradigm did not come without certain costs, however. In contrast to the flexibility and sensitivity of expression afforded by free phenomenological description (as practiced by William James, or the Gestalt psychologists, for example), the subject was now confined to the response categories that the experimenter made available, which were few in number (so as not to overload the subject with scoring information), and usually organized along a single perceptual dimension.

Even more significantly, perhaps, the procedure required the experimenter to postulate in advance which physical stimulus characteristics were psychologically relevant for the task at hand. Each possible response category was ultimately interpretable as a hypothesis about some public property of the stimulus. In a discrimination experiment, for example, the subject was not asked to state whether he "heard" a tone; what he had to assess, instead, is whether his degree of belief in a tone-presentation exceeded a certain criterion level. All experiences, subjectively tone-like or not, that supported sufficient belief would be covered by a common "tone" response, in this methodology.

The anchoring of the response categories to physical stimulus descriptions may not have been a problem in psychophysics, where the psychologically relevant stimulus properties are well understood, but it made it difficult to extend the signal-detection approach to exploration of mental processes that were not self-evidently implicated in the cognition of some simple objective variable. Knowledge of word-meanings, of facial

expression, of humor, of social convention, of perceptual similarity and grouping, could not fully benefit from the application of this method, because, in each case, the physical stimulus characteristics that engage the cognitive response appeared hopelessly obscure. The resulting division of substantive areas of cognitive psychology on purely methodological grounds goes against our intuition that recognizing, e.g., the ironic edge in a verbal remark is no different, subjectively, from detecting a weak physical signal in noise, notwithstanding that science will not anytime soon have an adequate physical description of 'ironic' stimuli.

Mutual Knowledge of Cognitive States

But, now, if we do not wish to condition the subjects' scores on public stimulus characteristics, we must then select some other kind of variable to take their place. What sort of variable might be appropriate for such a role?

Let us take a step back and consider that every introspective report, no matter how personal, is directed towards some audience. That audience may be large and ill-defined, such as the community of English speakers, it may be a small group of specific individuals, like a wine-tasting circle, or it may be a single person -- the experimenter in a psychological experiment, perhaps. But there always is an audience, and the role of that audience in interpreting the report must in some way be defined.

A wine-drawing by taster A, for example, will convey something to taster B only to the extent that it is a "piece" that fits into an existing representation that B has of A's perceptual possibilities. In the absence of any such representation, the drawing remains a cipher, meaningful as it may appear to A.

Ultimately, it may be possible to describe the knowledge demonstrated by someone who correctly interprets the curvature in a wine-drawing, or, the irony in a remark, as discrimination of some complex of physical stimulus attributes, but this, surely, is a very roundabout way of looking at the problem. Is it not much more natural to think of that person as discriminating directly certain psychological variables, occuring within the other person -- in this case, the author of the drawing/remark?

If we take this idea seriously, we are then led to conceptualize the knowledge that underwrites introspective reports -- that is, reports successfully generated and successfully received by the intended audience -- as a form of mutual knowledge of mental structure, and we are led to consider including the audience of these reports in the experimental procedure.

The ease with which we talk about our thoughts, sensations, emotions, does seem to imply that we can draw readily on a great deal of mutual knowledge of each other's mental life. Casual introspection, unregulated by any public criteria of communicative success and failure, hints at this knowledge, but does not prove its existence. As any participant at a wine tasting will confess, the discussion that takes place is tantalizingly inconclusive, even when some verbal consensus is ultimately reached.

The knowledge that one person has of another person's internal processes is no different in principle from knowledge he has of public aspects of his environment. Just as I may use my reaction to a given wine sample as a source of information about its physical properties, its sugar content, acidity, etc., so I may also use it to infer the likely attributes of another taster's experience of it, provided, in both cases, that I have an

appropriate representation of these other variables, and am able to read what my perceptual cues say about them.[2]

Experimental investigation of this knowledge is impeded however by one additional obstacle. If one wishes, for example, to assess a subject's knowledge of sugar content in wines, one can teach him, through appropriate feedback, how to discriminate sugar content as well as his sensory capacities allow. Now, in order to study, in a like manner, his knowledge of a variable, α, that characterizes the sensory experience of another person, the feedback necessary for teaching discrimination of this variable must, of course, be provided by the introspecting responses of the person within whose body α resides. But here one is confronted by an apparent circularity, since someone must first teach α's owner to consistently identify these values.

Introspective Scoring Rules

In this paper, we will investigate procedures that simultaneously elicit responses about a stimulus from several subjects, compare these reponses, and then score them competitively, in such a way that maximization of score requires each participant to fully identify the values of the relevant internal variables. Figure 2 shows the general schema for these procedures, which we will call introspective.

By comparing this to the signal-detection paradigm, shown in Figure 1, we see that things are different in several respects. First, although the experimenter selects and presents a stimulus to his subjects, he does not use his own information about the stimulus as a determinant of the

---

[2]Skinner (1945) contains an early but systematic discussion of these issues.

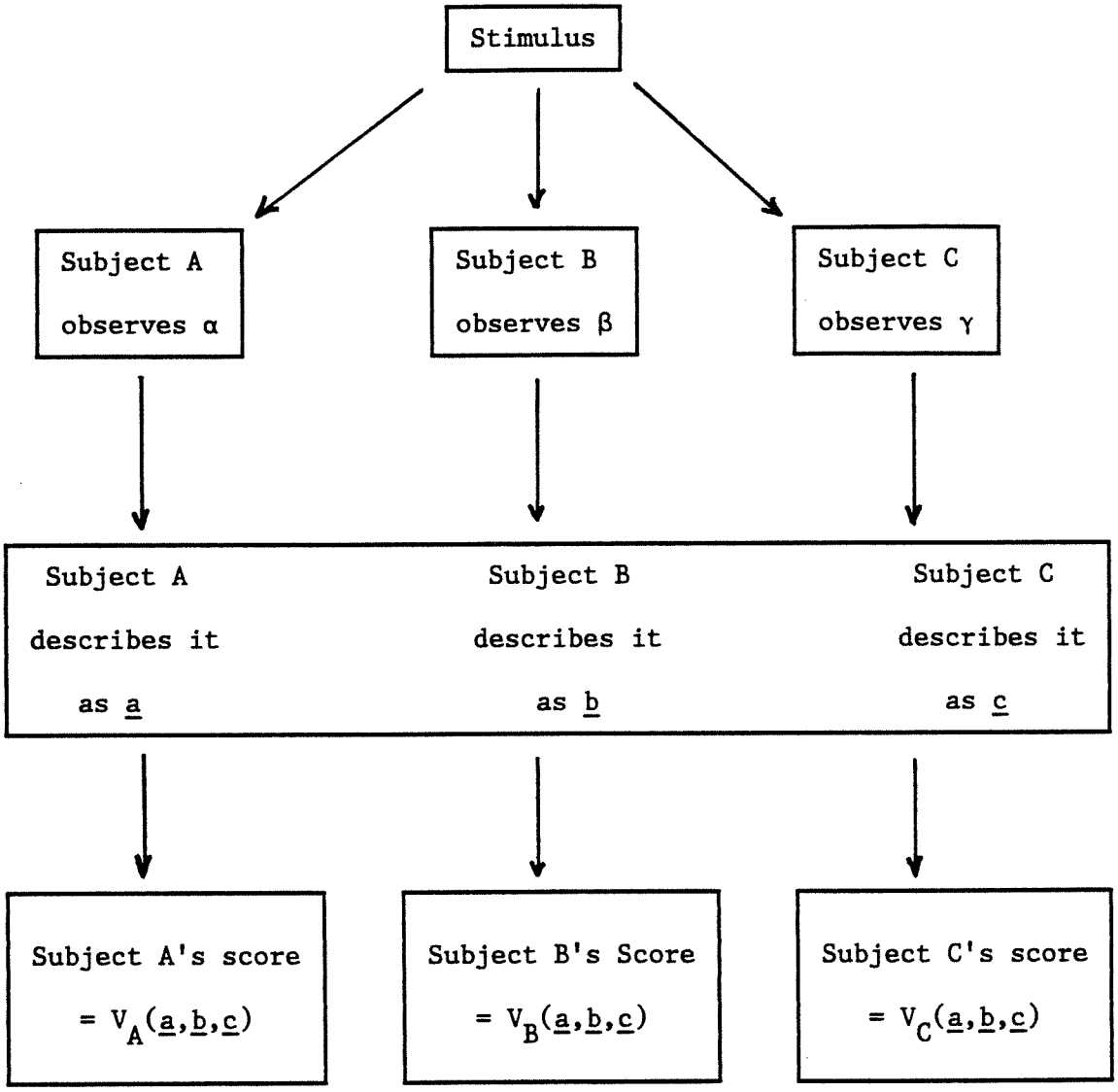


Figure 2: The introspecting paradigm: Each subject is instructed to maximize personal expected score under the assumption that other subjects are doing likewise.

Insert Figure 2 About Here

subjects' scores (the variable, s, does not appear in the schema). A key distinguishing feature of an introspective procedure, as I conceive it, is that the stimulus is methodologically treated as an unanalyzable unit, identified only by label. The experimenter may, of course have a great deal of information about the stimulus, and he will use this information when he analyzes and interprets the subjects' performance, but that knowledge does not have the privileged status of a criterion for correct responding. For example, the experimenter may feel strongly that blue is more similar to green than to red, but he would not wish to penalize subjects who disagree.

Second, every subject has a personal score, perhaps different from the scores of other subjects. If each subject is instructed to maximize his own score, under the assumption that the other subjects are attempting to do the same, then the entire procedure becomes in effect a game of strategy, and is amenable to analysis by the mathematical theory of games. This is a significant complication of the problem, and it is important to see why it is really necessary.

Were one to adopt the much simpler approach of assigning a common score, $V(a,b,c)$, to all participating subjects, then there would be no incentive for any one of them to make their response functionally dependent on the internal variables. There would exist, instead, some combination of responses, call them $a^\circ, b^\circ, c^\circ$, that maximizes the common score, and as the scoring function is public knowledge, there would be no reason for any subject to choose anything but the best response, irrespective of his or her assessment of the stimulus. A cooperative procedure, therefore, is generically incapable of differentiating between an empty verbal agreement,

and an agreement that reflects a genuine coordination of internal variables.

A third feature of the procedure, not represented in Figure 2, is the presence of one subject who does not observe the stimulus and whose score is inversely related to the scores of the other two. This subject plays the role of a "skeptical outsider," whose task is to simulate good introspective judgments, and so disprove that the others are revealing stimulus-relevant information.

A methodological criterion for introspective procedures

The objection to cooperative procedures that we have just raised, identifies a criterion that will be invoked repeatedly in this paper:

> Can subjects maximize individual expected scores in the procedure, without revealing values of the relevant internal variables?

In applying this test to a procedure, we will need to examine carefully whether clever subjects could score through misrepresentation or concealment of information. But, and here I would like to forestall a possible misunderstanding, this does not mean that we attribute to actual subjects similar powers of on-the-spot calculation of optimal strategies, nor that we impute to them any deliberate desire to withhold information; rather, we feel that when subjects participate in a group procedure that does not pass this test, then, just like the solitary introspecting observer, they will be vulnerable to falling into response patterns, i.e. "conversations," that

appear informative (to them), but that in fact are not so.[3] In this paper, the assumption of rational play is used as a formal device for identifying procedures in which this kind of mutual self-deception cannot arise.

Outline of paper

The next section introduces the basic ideas behind our solution, in the context of a very simple three-person game (Section 2). Section 3 presents a general method for modelling the mutual knowledge of internal variables that is created when several subjects observe a common stimulus. The main result of the paper, developed in Section 4, identifies a procedure that provides correct incentives for (ideal) subjects to reveal essentially all information about internal variables. Sections 5 and 6, which can be skipped on first reading, point out relationships to information theory (Garner, 1962; Shannon and Weaver, 1949), and work through examples of hypothetical play. Section 7 describes a sequential version of the procedure that, although more difficult to analyze, has greater experimental promise because it requires less work on the subjects' part. The paper concludes with a presentation of some illustrative experimental results, in Section 8, and a brief general summary in Section 9.

---

[3]Let me illustrate this with an example. We wish to test whether subjects can learn to classify a collection of stimuli in a coordinated way, and so instruct them to consider each stimulus, separately classify it, compare their responses, and go to the next one. A procedure in which subjects were scored by relative frequency of agreement would not pass our test, because it would pressure subjects to economize in categories (why not just one?), and even if there was a quota on the permissible proportions of different categories, the subjects could still manage to achieve coordination through inadvertent sequential patterns. In a difficult categorization task a subject simply may not know whether he is choosing a category because it is the most appropriate one, or because, not having chosen this category lately, its turn has come. It is precisely the subjects' lack of causal insight about their behavior (Nisbett and Wilson, 1977), that makes the application of this criterion important.

2 A GAME WITH PRIVATE CARDS

Since some readers may not be familiar with the game-theoretic analysis used here, it will prove useful to start with a somewhat more concrete version of the problem. Imagine, if you will, that you are charged with the task of designing rules for a game of cards, played like a regular game, but with one peculiar feature: At no point is any player required to place a card face up on the table; at no point, in other words, is he required, or in fact, allowed to prove that he holds a particular card. As in a regular game, cards are dealt face down to players seated around a table; the players examine their cards, make certain statements, then receive more cards, and so on. At the end of the game, the scores are tallied for each player, but these scores may be functions of statements only.

In what way does this task capture the essence of the problem stated in the introductory section? The card that is dealt to a player represents his subjective assessment of a stimulus, and like the assessment, it is private, and cannot be shown to others. But, just as one's personal reactions to a stimulus allow conjectures about possible reactions of another person, so, too, the cards that one holds in the game say something about what cards might be held by other players. Even if the deck is shuffled before dealing, I still know, for example, that if I hold the Ace of Spades, then no one else can hold the Ace of Spades. The dealing of the cards creates a web of mutual inference, which it may be possible to make strategically useful through a suitable choice of rules. To the extent that we can invent rules that make this a meaningful game, that is, a game in which successful play requires players to attend to their cards, and make statements according to them, to that extent we have created a set of

instructions for rewarding identification of variables that are directly accessible to only one person.

A Three-Person Game of Matching Pennies

Here is an example of a three-person game that works in the way we just specified, and which will serve as a vehicle for introducing the basic properties of the more elaborate procedure we will develop later on. To keep things simple, we reduce the standard 52-card deck to only four cards, by selecting one representative from each of the four suits.[4] The game is played by three players, of whom one, the Dealer, begins by distributing two cards to the player on the left (called Player A), and a single card to the player on the right (called Player B). The remaining card is left face down on the table. The Dealer holds no cards. Play consists of a single round of "bidding," in which each of the three players privately writes down one of two statements, either "Heads" or a "Tails". Should all three statements agree, the Dealer loses some fixed amount to the two other players; in all other cases, i.e. when statements do not agree, the two players pass that amount to the dealer. The game is played only once; public discussion is allowed, but pairs of players do not have any private means of communication. We can think of this as an elaboration of the two-person Matching-Pennies game, where two persons simultaneously declare Heads or Tails, with one winning if the two choices agree, and the other if they disagree (see Figure 3).

Insert Figure 3 About Here

---

[4] Namely, two red suits: Diamonds and Hearts; and two black suits: Clubs and Spades.

Figure 3: The Three-Person Matching-Pennies Game. Player A chooses Row, Player B column, and the Dealer chooses Left or Right Matrix. The entries indicate the winning side.

Strategy 1: Randomization

Thinking informally about this game, we see that the two partners, A and B, must create some unpredictability about their statements in the Dealer's mind, because of they agree publicly to go one way or the other, then the Dealer will always be able to defeat a match. Let us see first how well they can do if they do not attend to the cards. Lacking a private means of communication, they can do no better than to randomize--to each privately flip a coin--and play accordingly. In that case, the Dealer will also randomize, to protect himself against being outguessed by his two opponents. The probability of a three-way match will be 1/8 plus 1/8, or 1/4. Without cards, then, the two partners cannot prevent the Dealer from winning with at least 3/4 probability.

Strategy 2: B leads

The perceptive reader may have noticed, however, that the cards are of some use here. Let us suppose that the Player A receives a red card -- a Heart or a Diamond. Since this leaves only one red card in the deck, black cards are going to have a higher chance of appearing in the other player's hand. Consequently, by announcing publicly that he will play Heads if and only if he holds a red suit, Player B may hope to reduce the uncertainty about his statement in his partner's mind, while simultaneously keeping the Dealer in the dark.

Let us now follow through the implication of this announced strategy, from the perspective of Player A. The two cards that he holds then fall into three strategically distinct types:

1. With probability 1/6 he holds two black cards, in which case he should play Heads, since B holds a red card, and will play Heads.
2. With probability 1/6 he holds two red cards, in which case he should play Tails, since B holds a black card, and will play Tails.
3. With probability 2/3 he holds a red and a black card, in which case he has no knowledge of B's card, and can choose Heads or Tails at random.

The probability that A and B will agree is thus 1/6+1/6+(1/2)(2/3), or 2/3. The Dealer still picks at random (since the probabilities of two-way matches on Heads and Tails are the same), thus cutting down the probability of a three-way match to 1/3. For A and B, this compares favorably with the 1/4 probability of winning that they face if they ignore their cards and randomize.

Strategy 3: A Leads

Is there an even better strategy? Suppose A reasons like this: He finds in his hand a Spade-Heart, say, so his partner must hold one of the remaining cards, a Club or a Diamond. A cannot tell which one of the two is held by B, but in both cases he knows that B, in reasoning about A's cards, can exclude the Club-Diamond pair as a possibility. Thus, if A announces that he holds either Spade-Heart or Club-Diamond, and that he will play Heads for the former, and Tails for the latter, then he has communicated his bid to B without leaking any relevant information to the Dealer. With this strategy, the game becomes a fair one, as A and B cannot fail to match, and will win with probability 1/2, at least. No further improvement is possible, since the Dealer can prevent a full match one-half the time by flipping a coin. (Notice, incidentally, the importance of allowing communication throughout the game; if communication is permitted

only prior to the dealing of cards, then Strategy 2 (or some permutation of it) is the best that A and B can do.)

Remarks About the Game

We have here a game in which successful play requires certain players to reveal aspects of private information, even though the scoring rule only evaluates actions. Any form of inconsistency on the part of A and B -- either not attending to the cards, or not following the agreed convention for translating them into bids -- will lower their chances of winning.[5]

---

[5]Are three players necessary? There do exist two-person games that elicit private information, but they are strategically more complex. Here is an example, adapted from Aumann (1974), that illustrates the problem: A and B each draw a card from a three-card deck, containing a Blue, Green, and a Red card. Subsequently, they each privately select one of three bids (1, 2, or 3); after the bids are compared, Player A receives the first, and B the second entry indicated by the corresponding box in the matrix (negative entry means loss):

	b_1	b_2	b_3
a_1	-2,-2	2,0	0,2
a_2	0,2	-2,-2	2,0
a_3	2,0	0,2	-2,-2

The game is symmetrical, so we might as well consider how it appears to A. If he prematurely reveals his bid, (say, a_1), then B will counter with the bid (in this case b_3) that gives A nothing (we assume that they are prevented from sharing). If A and B do not communicate at all, and bid randomly, their expected score is again zero. However, if A announces that he will bid 1, 2, or 3, depending on whether he holds the Blue, Green or Red card, he then invites B to adopt the same contingent strategy, since only in that way can the negative diagonal entries be avoided. In this correlated equilibrium (Aumann 1974; 1987) the expected score of each player is +1. The game combines cooperative, and competitive elements. We suspect that even in the absence of cards, players who play this game repeatedly could arrange to share the benefits of coordination by cycling through the non-diagonal entries in some predictable way.

Although the example shows that it is possible to elicit private information through games of strategy, this particular game is not a good instrument for the purpose. First, by giving each subject the same response set -- Heads/Tails -- and scoring the game for matches, we have defined the problem as one of agreement rather than communication. Insightful players may indeed discover how to translate their observations into elements of the response set, but the process of translation, of figuring out the best way to bid as a function of cards, is intellectually taxing and draws attention away from the real purpose of the procedure, which is to obtain a true record of each player's cards.

Second, the game works because of a fortuitous match between the scoring and the probabilistic inter-relationship between the cards held by A and B. The scoring rule, in other words, was designed in full knowledge of the process by which the hands dealt to A and B were generated. A change in the composition of the deck, or a change in the number of cards dealt to A and B, would require one to modify the rules of play, if one wished to obtain comparable information. But in any less artificial example, the person who designs the rules will not know how the observations of different players are inter-related, and will need a procedure that works in the absence of such knowledge.

3 REPRESENTATION OF MUTUAL KNOWLEDGE

The structure of mutual knowledge in the Matching-Pennies game

The problem before us, then, is to find an elaboration of this game, that:

(a) is strategically transparent,

(b) requires each player to identify his or her observation,

(c) makes few specific assumptions about the way in which the observations of different players are co-determined.

The theory of Bayesian-Games (Harsanyi, 1967) is a multi-person extension of statistical decision theory (Luce and Raiffa, 1957) that provides a framework within which this problem can be clearly formulated, and solved. The most important aspect of the theory, for our purposes, is that it gives us a general way of representing mutual (but imperfect) knowledge of internal variables. Here is how this knowlege would be modelled for the game we just described.

The information available to Player A, denoted by α, are the cards in his hand:

$$\alpha \in \{SC, SH, SD, CH, CD, HD\}.$$

The information available to Player B, which is denoted by β, is a single card:

$$\beta \in \{S, H, D, C\},$$

Player C holds no cards, so his information, γ, is an element of a singleton set, which we need not write down.

If the cards are shuffled before dealing, then the joint probability of any combination, (α,β), appearing in the hands of A and B, is either zero or 1/12, depending on whether or not α and β are a possible combination (see Table 1). This joint probability distribution we will write as,

$$p(\alpha,\beta) = \Pr\{\text{A holds } \alpha \text{ and B holds } \beta\}.$$

To hold notation to a minimum, we will use the same symbols for random variables, $\alpha,\beta\ldots$, and the generic values of these variables. Specific values, however, will be starred or subscripted.

If players are able to calculate $p(\alpha,\beta)$, then the inferences that any player draws from his hand are derived by conditioning the joint distribution in Table 1. When A is dealt α^*, for example, then his beliefs about B's cards are given by the conditional probability distribution,

$$p(\beta|\alpha^*) = \Pr\{\text{B holds } \beta \text{ if A holds } \alpha^*\},$$

(which can be calculated directly by applying Bayes' rule: $p(\alpha|\beta)=p(\alpha,\beta)/p(\beta)$), while B's beliefs, should he be dealt β^*, are given by:

$$p(\alpha|\beta^*) = \Pr\{\text{A holds } \alpha \text{ if B holds } \beta^*\}.$$

The dealer's beliefs, since he does not hold any cards, are given by the joint distribution, $p(\alpha,\beta)$.

A convenient visual representation of the information structure may be obtained by writing down the set of all possible deals from the deck, and then, for each player, partitioning that set into subsets whose elements

	C	H	D	S	
DH	1/12	0	0	1/12	1/6
SC	0	1/12	1/12	0	1/6
HC	0	0	1/12	1/12	1/6
SD	1/12	1/12	0	0	1/6
HS	1/12	0	1/12	0	1/6
DC	0	1/12	0	1/12	1/6
	1/4	1/4	1/4	1/4	

Table 1: $p(\alpha,\beta)$ for Matching-Pennies Game

are all the deals that the player considers possible, when he receives a particular hand. In our example there are 12 distinct deals, and they are displayed in Figure 4, together with the relevant partitions. Figure 5 shows how the partitions are refined after Player A announces that he holds one of two complementary card-pairs. We will designate subsets of observa-

Insert Table 1 About Here

Insert Figures 4 and 5 About Here

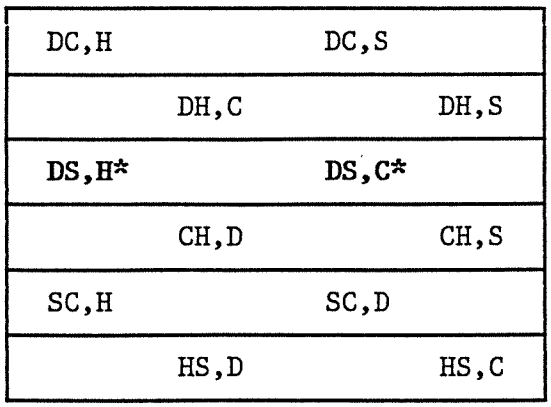
tions by enclosing the observational variable within parentheses, and, when appropriate, subscripting this with a property or aspect that identifies the subset, e.g.:

$$(\beta)_{RED} = \{H,D\}.$$

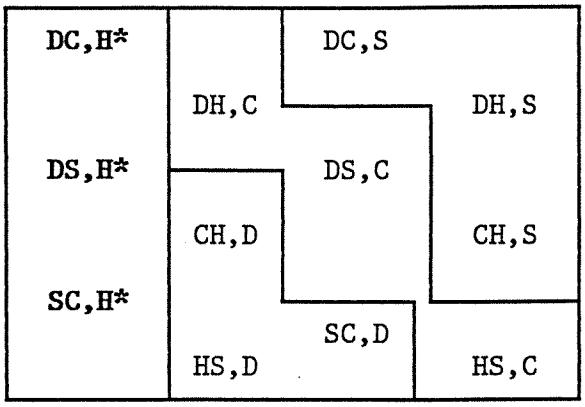
What is an observation?

The cards in this example were introduced as conceptual placeholders for a person's subjective reaction to a stimulus. We now turn our attention to situations where mutual knowledge, instead of being generated by a simple statistical mechanism (i.e. card-dealing), arises from the co-ordinated presentation of some experimental materials.

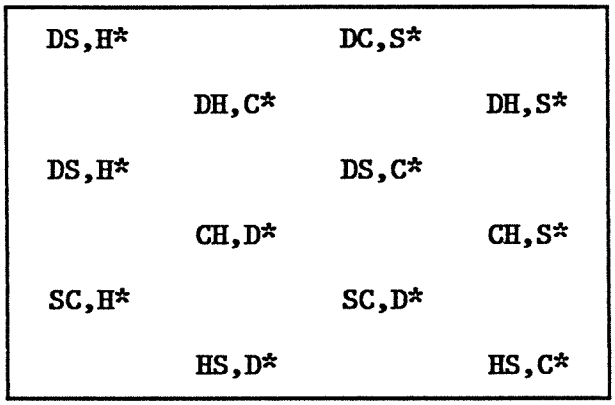
The setting for the introspective task, therefore, is a specific experimental episode, in which several participating subjects (we shall limit ourselves to three) are privately presented with some (not necessarily identical) matter for evaluation, e.g. a perceptual stimulus, a text, or a problem. As a result of this observation period, each subject is prompted into some form of cognitive activity. He may, for example, recognize the



Player A

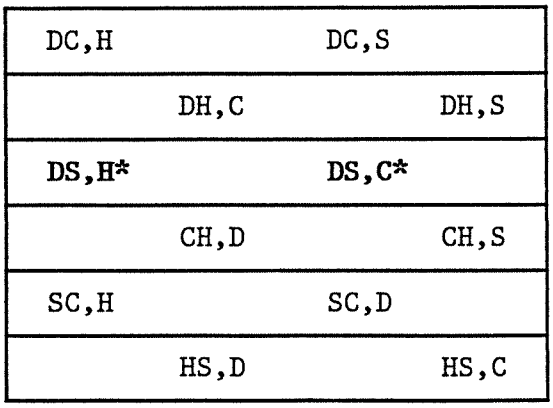


Player B

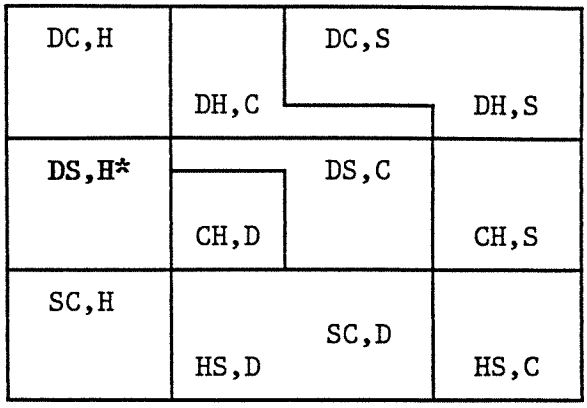


Dealer

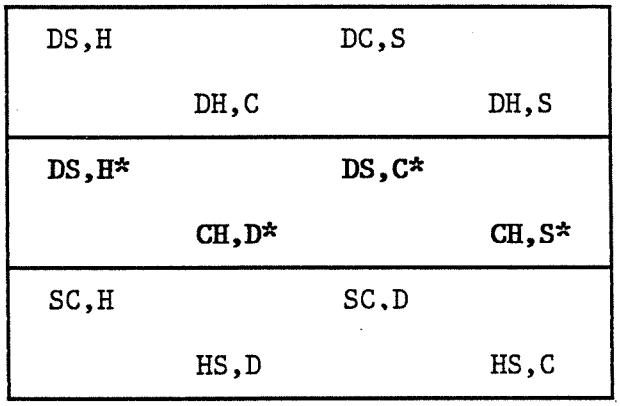
Figure 4: The initial partitions of the three players. The actual deal is Diamond-Spade to Player A, and Heart to Player B. Stars identify the deals that each player considers possible.



Player A



Player B



Dealer

Figure 5: The partitions after Player A announces that he holds either a Diamond-Spade or a Club-Heart hand, and the other two players believe him. Player B, in particular, knows now what the actual deal is.

stimulus as being of a certain type or category, notice some of its features, understand its significance, feel pleased or shocked by it, remember that it was shown once before, and so on. Ignoring these and numerous other distinctions, I will say simply that the subject has observed the stimulus, or, alternatively, that he or she has made an observation of it. I will not, in particular, differentiate between observations that would normally be interpreted as statements about the stimulus, from those that would be interpreted as phenomenological reports of feelings, sensations, and the like.

Although in discussing observations I will have to make use of whatever terms ordinary language offers, I wish to concept to be understood rather as a sort of phenomenal snapshot, capturing all aspects of experience that an individual can distinguish, irrespective of his current ability to describe these aspects. Hence, one should think of an observation as the ideal, most informative account of mental activity that an individual is capable of producing, were he instructed in the appropriate vocabulary by an omniscient experimenter.

In this paper, we will not specify the internal structure of the set of possible observations in any detailed way; however, occasionally it may be useful to think of the set as generated by some array of variables. If the objects under study are thought to be represented dimensionally, then an observation would be picked out by a vector of real numbers, identifying the levels on the dimensions at which this observation falls. Alternatively, we could think of specifying an observation by interrogating a long list of binary aspects, or features, and checking off those that do apply. Following Tversky and Sattath (1979), we would allow a feature "...to describe any property, characteristic, or aspect of objects that are relevant to the task under study," as well as adding here that features

need not be restricted to public aspects of objects, but may especially cover aspects that are only noticed by a single individual. For this reason, we would prefer to think of each subject as having a personal inventory of features, from which his observations are then determined. Should the features of different subjects align in such a way to create a common ground of agreement among them, all the better, but such agreement is not a necessary starting point for our analysis.

Mutual knowledge in the introspecting experiment

After observing the stimulus, each subject is in a position to formulate conjectures and hypotheses about other participating subjects, who, as we assumed, were shown the same or related stimuli. One person's observation generally carries in train a multitude of beliefs about what other people might have observed, that is, what they might have noticed, recognized, felt, thought about, etc. during the observation period. These beliefs may go no further than to ascribe to others exactly what the person himself has observed; or, again, they may attempt to compensate for differences in perspective, attitude, or discriminative ability.

We assume that each subject is able to systematize, and express his beliefs about other subjects' observations, by assigning a subjective probability to the event that the other subjects' observations are characterized by some aspect or property, which is to say, that these observations fall into the subset of possible observations sharing this property. We assume, also, that these subjective probabilities satisfy the normal rules of probability calculus (for unions, complements, etc.).

Having gone this far, we must also allow each subject to have well-defined beliefs about what someone else might believe about his own

observation. This second generation of beliefs then gives rise to a third generation of beliefs about beliefs about beliefs, and so on and on (Lewis, 1969). In what follows, we seek relief from this complexity by assuming that,

Condition 1 The beliefs of each subject are given by one of finitely many distinct subjective probability distributions over other subjects' possible observations.

This assumption partitions the space of all possible observations into finitely many equivalence classes, so that observations within one equivalence class are associated with the same beliefs (and beliefs about beliefs...) about other people's observations.[6]

Having endowed each subject with a fully developed set of beliefs about the observational possibilities of other individuals, we now address the question of how these beliefs might be coordinated. The assumption that we will presently make is that the beliefs of all subjects can be derived from a probability distribution over a structure of overlapping partitions, like the one we introduced in modelling the Matching-Pennies game. Game-theoretic literature refers to his as the assumption of consistency (Harsanyi, 1967):

---

[6]The assumption of finiteness is not conceptually restrictive. It is well known, for example, that any dimensional representation can be approximated by a finite number of features through a sufficiently fine partitioning of the continuum (see, for example, Tversky and Sattath, 1979). It is also possible to show that much more complex mathematical structures, such as those needed to rigorously represent the infinite hierarchy of beliefs (i.e., probability distributions) about beliefs, may be approximated to an arbitrary degree by finite structures (Mertens and Zamir, 1985, Section 3).

Condition 2 (Consistency) The beliefs of each subject, A,B,C, are derived by conditioning a common probability distribution, $p(\alpha,\beta,\gamma)$, on that subject's actual observation, $\alpha^*,\beta^*,\gamma^*$.

For example, if Subject A observes α^*, then the probability that he assignes to the event that B's observation is characterized by aspect (β) is,

$$\sum_{\beta\varepsilon(\beta)} p(\beta|\alpha^*).$$

In the context of the Matching-Pennies game, the meaning of this assumption is clear enough: it states that all subjects understand the statistical properties of the process by which they arrive at their observation. These properties, in the example, can be calculated by counting the possible ways in which cards can be dealt from the deck.

The meaning of this assumption in our experimental procedure is somewhat more difficult to grasp. The assumption says that subjects have a common understanding of the uncertainty in the natural process by which they arrive at their respective observation, and in this respect it is the natural generalization of the ideal-observer assumption, derived from signal-detection theory (Green and Swets, 1966). The assumption also implies that although subjects may know little about each other's observations per se, they are nonetheless in complete agreement about the implications of any particular observation by one subject for the observations of

others. One could say that the informational content of every possible observation is common knowledge among them.[7]

Two forms of strategically irrelevant information

The list of observations that we drew up in the Matching-Pennies Game was very selective (being just an enumeration of possible hands), and excluded many additional aspects that might have been noticed by some, or all players. For example

(1) B notices a pencil mark on the face of his card;

(2) The backs of cards are blue.

Example (1) picks out a distinction that is strategically useless in any communication game, because it does not affect B's beliefs about the observations of the other players: the presence of the mark is uncorrelated with any other noticable event in the situation. Recal that in an introspecting procedure (viz. Figure 2), a person's observations are useful to him only to the extent that they predict something about the actions of other players'; these actions, in turn, are functions of their observations, so aspects of observations that have no predictive power about another player's observations will a fortiori have no predictive power about his actions, and will hence have no effect on play. For this reason, it might be more proper to speak of equivalence classes of observations,

---

[7] In further defense of this assumption, game theorists point out that whenever it is possible that different players hold different theories about the probabilities governing Nature's toss, one can take a step back and include this uncertainty about other players' theories in the preliminary canvassing of all possible observations (Aumann, 1974; 1987). Extending this to our context, we could say that from a subject's perspective, Nature not only selects a stimulus, but also the characteristics of the people brought together in the experiment, and a part of these characteristics are their personal theories and hypotheses about how people typically react to these stimuli.

where a single equivalence class would cover observations with a common belief-signature, i.e., having the same implications about observations of other subjects. Since these observations are in turn individuated in exactly the same way, the entire structure is one of mutual interlock of belief.

Example (2) defines an aspect that is public in a very strong sense, for not only does everyone know this is true, but everyone knows that everyone knows that it is true, ... and so on. It is customary to call such aspects common knowledge (Aumann, 1974; Lewis, 1969; Milgrom, 1981), and they, too, are irrelevant to play in communication games: The possibilities of different players engaging in strategic coordination is not affected by things that everyone knows (such as the names of the players, the time of day, facts of arithmetic, etc..).

The need for an uninformed player

The irrelevance of common knowledge presents a problem for the design of a communication game. When, for example, we show an object to the participating subjects, we know that they will have no reason to reveal aspects of this object that are common knowledge among them, and these aspects, after all, may be the most salient or interesting ones. To get around this fundamental limitation on communication games -- that they can only elicit information that is unevenly distributed among the players -- we will generalize a feature of the Matching-Pennies Game, and postulate that one of the subjects in the procedure does not have access to the stimulus:

Condition 3 There is one uninformed subject, C, whose beliefs about the other subjects' observations are not affected by his or her observation.

How ought one to interpret the beliefs of this uninformed player? They are whatever can be gleaned from the public aspects of the experiment: from the common initial instructions, or from previous experience as an uninformed player.[8]

A wine-tasting example

The model we are finally left with is simple but flexible. Two informed subjects each draw an observation (α^* and β^*) from a distribution $p(\alpha,\beta)$, which also summarizes the information available to the uninformed subject. All that A and B know each other's observation is then contained in the two conditional distributions $p(\beta|\alpha^*)$ and $p(\alpha|\beta^*)$. Clearly, we have here a modelling framework that makes as few as possible specific assumptions about the psychological mechanisms that mediate individuals' knowledge of internal states, and of their ability to communicate them. A particularly glaring omission is the absence of any account of attentional constraints, that is, of how the process of describing some aspects of experience may limit awareness of others.

Having acknowledged these limitations, let us turn to a semi-realistic example of how the model might be applied. Suppose that the two informed subjects are about to taste a wine. Each subject can describe a wine with

---

[8]An alternative way of creating artificial informational asymmetries would have been to give each subject partial access to the stimulus; that approach, however, is conceptually less direct and is more difficult to implement experimentally.

three predicates, which yields a total of eight distinct observations (for each subject). Observation α_{101}, for example, indicates that A would describe the particular wine as having a Yes-value on the first and third aspects, and a No-value on the second aspect. Comparing the partitions in Figure 6, we see that the two subjects' aspects are somewhat co-ordinated, so that aspect (α_1), of A corresponds to $(\beta)_1$ of B, and so on. Specifically, on the first two aspects, the difference between A and B is one of

Insert Figure 6 About Here

selectivity: The wines that get a Yes-designation on $(\alpha)_1$ are a subset of those that get a Yes-designation on $(\beta)_1$, and the same relationship holds between $(\beta)_2$ and $(\alpha)_2$. (We might say that A has a stricter criterion on the first aspect, and B on the second.) The final pair of aspects, $(\alpha)_3$ and $(\beta)_3$ are also related, but not in such a simple way.

The probability distribution for all 64 possible observation-pairs can be read directly from an overlay of the two partitions, assuming that we have drawn the partitions so that the probability of an event is proportional to area. This distribution is given in the matrix below.

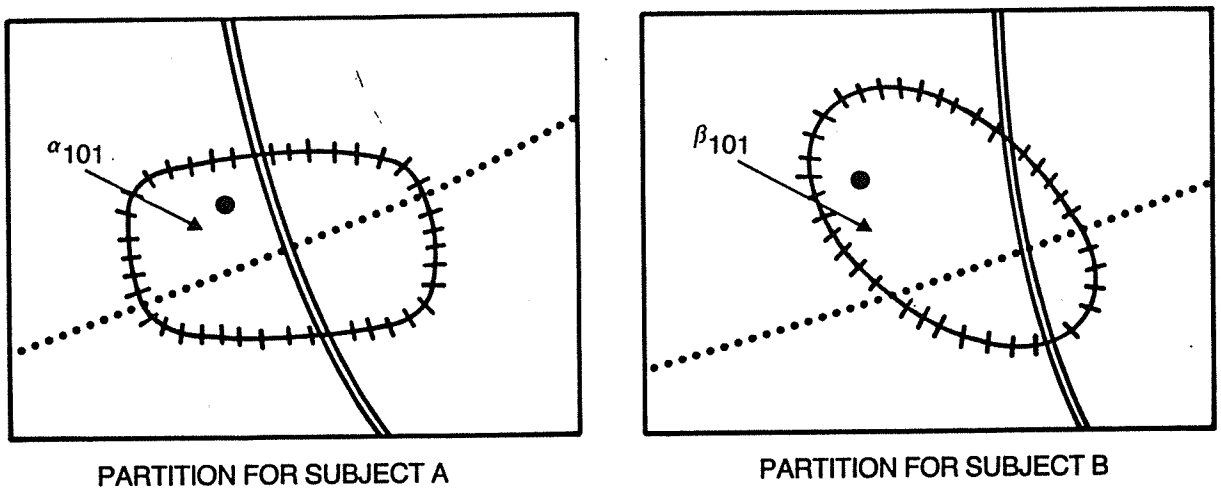


Figure 6: The dot identifies what happened on this trial.
A's observation, α_{101}, is characterized by aspects $(\alpha)_1$ and $(\alpha)_3$;
B's observation, β_{101}, is characterized by aspects $(\beta)_1$ and $(\beta)_3$.

	β_{000}	β_{001}	β_{010}	β_{011}	β_{100}	β_{101}*	β_{110}	β_{111}	
α_{000}	.087	.002	.000	.000	.049	.018	.000	.000	.156
α_{001}	.002	.006	.000	.000	.000	.044	.000	.000	.052
α_{010}	.049	.000	.116	.010	.000	.000	.018	.006	.199
α_{011}	.000	.016	.000	.010	.000	.028	.000	.022	.076
α_{100}	.000	.000	.000	.000	.207	.012	.000	.000	.219
α_{101}*	.000	.000	.000	.000	.051	.043	.000	.000	.094
α_{110}	.000	.000	.000	.000	.028	.000	.128	.000	.156
α_{111}	.000	.000	.000	.000	.020	.014	.006	.008	.048
	.138	.024	.116	.020	.355	.159	.152	.036	

It is important to realize that this matrix contains information about all possible conjectures that A, B, and C can make concerning each other's observations. Suppose, for example, that on this occasion the location of the wine is given by the dot in Figure 6, so that the wine is rated α_{101} and β_{101} (by A and B) . From his observation, A can can narrow down B's observation to one of only two possibilities: β_{100} and β_{101}. B, however, is much more uncertain about A's observations; the only ones he can exclude are α_{010} and α_{110}. In particular, the probability that A assigns to B's actual observation, $p(\beta_{101}|\alpha_{101}) = .46$, is greater than the probability that B assign to A's actual observation, $p(\alpha_{101}|\beta_{101}) = .27$.

We can go one step further, and calculate second-order beliefs. What does A think that B thinks is the most likely of A's observations? It is neither the actual observation, α_{101}, nor the observation that B <u>actually</u> thinks is most likely, α_{001}, but yet a third observation, α_{100}. The expected probability (in A's mind) that B assigns to this α_{100} is,

$$p(\alpha_{100}|\beta_{100})p(\beta_{100}|\alpha_{101}) + p(\alpha_{100}|\beta_{101})p(\beta_{101}|\alpha_{101}),$$

or .35; which is greater than the expected probability that B assigns to the actual α_{101} (.20). This shows, incidentally, that in a game in which A and B were trying to match descriptions, and in which B was indeed correctly describing his observations, A would have reason to mislabel α_{101} as α_{100} (so inviting mislabelings on the part of B as well).

4 THE BASIC COMMUNICATION GAME

Two-on-One Games

The procedure thus begins with a brief observational episode, that two persons share, but from which a third, otherwise identically situated individual is excluded. This creates an asymmetric structure of mutual knowledge in which the two informed participants have less uncertainty than the uninformed person about the aspects of each other's observations.

A class of games in which this differential mutual knowledge has an essential function in rational play, are games in which two players are joined in an alliance against a common opponent. The Matching-Pennies Game is an example of this class, which we shall call Two-on-One Games.

Definition 1 A Two-on-One Game is a three-person game, in which players may be labeled A, B, and C, so that their respective scores, V_A, V_B, and V_C, satisfy the following two conditions:

(1) $V_A + V_B + V_C = 0$;

(2) $V_A = V_B$.

The scores have only one degree of freedom, so that we can without ambiguity speak of V_A as the score for the game, and, dropping the subscript, write it as V. In any Two-on-One game, the goal of A and B is to maximize, and of C to minimize the score.

The rules

We would like now to specify the rules for a Two-on-One Game in which rational play by A and B will require them to fully explain, through their actions, what they have observed. Embedded within the available actions there must exist some that may be interpreted as the definition of a set of possible observations, and, the selection of one element -- the "actual observation" -- from this set, as well as some that correspond to the labelling of the actual observation by its associated probability distribution over the other player's observations. The question, now, is how to insert C's actions into the game, and then score the entire combination in such a way that A and B are penalized by a lower expected score whenever they commit one of the three introspecting mistakes, namely:

(1) omit possible observations from the set;

(2) incorrectly identify the actual observation;

(3) incorrectly report the beliefs engendered by this observation.

The second sort of mistake covers not only systematic misreporting of observations, but also simple inconsistency, -- i.e. the use of different labels for the same observation. We would like both of these mistakes to be penalized by a reduced score.

To keep the rules and notation as simple as possible, we will describe an asymmetric game, in which only A is asked to describe his observation.

The game can be made symmetric by requiring subjects to simultaneously play two such games, with the roles of A and B reversed.

There are two parts to the game, both of which take place after the observation period. The first part consists of a public conference, at the end of which A defines a personal vocabulary set, whose elements -- or descriptions -- are possible accounts of what took place. Individual descriptions may refer to subjective experience, or to public properties of the object; they may be generated by listing attributes, features, levels on a rating sale, and so on.

The entire set of possible descriptions can be presented by exhaustively listing all of its elements, or it could be defined by means of a set of rules for constructing possible descriptions. In the latter case, we shall speak, alternatively, of A's descriptive system, rather than vocabulary. Examples of acceptable vocabularies/descriptive systems would be:

(1) Sequences of no more than N letters from the English alphabet;

(2) Yes/No values for a set of binary attributes;

(3) A set of mutually exclusive English statements.

(4) Locations of a point in a two-dimensional array.

(5) Expressions formed by a discrete notational system, such as music scores.

Clearly, A has a lot of flexibility. What is important is that the vocabulary be discrete -- that it is always obvious whether two descriptions are equivalent (Goodman, 19xx). A pictorial system, like the wine-drawings developed by Vandyke Price, would have to be converted into a discrete format for this game.

The first part concludes when A privately selects a single description from his vocabulary. In the second part, B and C each separately assess a probability distribution over elements in the vocabulary set.[9] A's description is then made public, and the round scored in the way described below.

Definition 2 The Basic Communication Game is a Two-on-One Game in which Player A presents a vocabulary, A, and privately selects a description, a*, from it, after which Players B and C assess probability distributions, b(a), c(a), over possible descriptions, a ε A, and the score calculated as:

$$V(a^*, b(a), c(a)) = \log\left(\frac{b(a^*)}{c(a^*)}\right) .$$

Numerical scores will be computed with the base-two logarithm. For example, if the probability assigned to the correct description by B is twice as great as that assigned by C, the score would be +1; if it was four times higher, the score would equal +2, and so on.

Interpretation of the rules

Before analyzing the game formally, let us first try to develop some intuitions about what might constitute rational play. The score as we have defined it measures, in a particular way, the accuracy of B's guess rela tive to that of the uninformed player, C:

$$V = \log b(a) - \log c(a).$$

---

[9] We ignore here the practical difficulty of eliciting distributions over large sets of events. Sections 7 and 8 describe a procedure that gets around this problem.

Since the logarithm is an increasing function, and since B wishes to maximize, and C to minimize the score, it is clear that each of the two will bet a greater portion of their unit probability total on those descriptions that appear more likely to have been selected by A. But this observation would have been valid if we had substituted any increasing function for the log.

The decision to measure the ex-post accuracy of B's and C's guesses along a logarithmic scale is dictated by a remarkable (and for our purposes critical) property of the scale, which is that it is the only one that provides no incentive for persons to misrepresent their true subjective probabilities. We will derive this property shortly, but for now let us just make note of it, and return to Player A, who, in view of this, knows that B's and C's assessments will accurately reflect the probabilities they assign to various descriptions.

The problem for A, then, is to present a vocabulary that shapes the expectations of B and C in such a way that the differential accuracy of B's assessment against that of C is maximized. What sort of considerations will affect the design of an appropriate vocabulary? Clearly, A should allow for a description of the actual observation, but, in order to create as much doubt as possible in the mind of C, he will also want to surround the correct description with many plausible alternatives. The presence of such alternatives will force C to spread a finite probability mass over a greater number of descriptions, and thus decrease, on average, C's expressed degree of belief in any single one. These alternatives, at the same time, must not mislead B (too much). So, when in doubt whether to include a description in the vocabulary, A must consider whether the description is more likely to mislead B or C. Descriptions that obviously

cannot apply will not fool C and their inclusion has no effect on subsequent play.

Let us try to imagine what the preliminary discussion in Part 1 is like. It is, first of all, an unusually constrained discussion, because A cannot identify the description that he intends to select, nor can he say anything about what has transpired in the observation period. Remember that C is present at this conference, so that any information revealed here immediately becomes common knowledge, and is thereby made useless for the purpose of coordinating reports in Part 2.

Indeed, the only useful statements that A can make are conditionals, of the form, "If ... is true, then I will choose description $\underline{a}$" where the dots are filled with some statement about the observed object. Now, since in our formalization a person's observation fully determines all possible evaluations of the object, the class of relevant statements can actually be narrowed further to those of the form, "If I observe α then I will chose description $\underline{a}$ with a certain probability." All discussion, therefore, can be summarized by a conditional probability matrix that indicates the probability of selecting any description, conditional on observations.

It need not be necessary for A to fully specify such a matrix, if he feels that the configuration of descriptions in the vocabulary implicitly defines a list of conditional statements. For example, in the absence of any special qualifications, a vocabulary consisting of color terms only, declares, in effect, that the standard conventions for using color terms apply, i.e., "If the object is red, I will play 'red'," and so on. However, A is permitted to legislate new conventions, such as: "If the object is red, I will play 'red' with probability 2/3, and 'blue' with probability 1/3," if he feels this will work to his advantage. In any case, we assume that once the vocabulary is set, the convention automatically becomes

common knowledge, whether by accepted standards of usage, or through explicit specification of the relevant conditional probabilities.

The divergence of two distributions

When a logarithmically-scored assessor reports a probability distribution, $x = (x_1,\ldots,x_n)$, over a set of events (indexed by $\underline{i}$), his subjective expected score (evaluated, of course, before he learns of the true event, $\underline{i}^*$, by his true subjective probabilities, which we denote by y:

$$\text{Expected score} = \sum_{\underline{i}} y_{\underline{i}} \log x_{\underline{i}}. \qquad (*)$$

In the previous section, we alluded to a special property of this logarithmic scoring rule:

Proposition 0 In order to maximize subjective expected score along a logarithmic scale of accuracy, the assessor should match exactly the reported probabilities to the subjectively perceived ones. If the elicited probability distribution is defined over three or more events, then only the logarithmic scale has this property.

Checking this result (attributed to Andrew Gleason; cf. Savage (1971)) is not hard: The marginal benefit in subjective expectation due to a marginal increase in probability reported for event $\underline{i}$, is given by the derivative of (*) with respect to $x_{\underline{i}}$:

$$\frac{\partial(*)}{\partial x_{\underline{i}}} = \frac{y_{\underline{i}}}{x_{\underline{i}}}.$$

The optimal reported probabilities, $x_{\underline{i}}^{\circ}$, equalize the marginal returns across all events $\underline{i}$, which means, then, that truth-telling is indeed

optimal policy: $x_i^\circ = y_i$.[10]

One could restate Proposition 0 by saying that a person incurs a net loss in expected score whenever he misreports his subjective probabilities. Specifically, the expected loss, on average, is the difference of two expectations, one evaluated with true and one with false reports:

$$\text{Expected loss in score} = \sum_i y_i \log y_i - \sum_i y_i \log x_i.$$

This quantity will play a central role in the derivations that follow, so we will give it a special name, as the divergence of one distribution from another:

Definition 3 The divergence D(x,y) of probability distribution x from y, is the expected loss in logarithmic score due to reporting distribution x, when the true subjective distribution is y:

$$D(x,y) = \sum_i y_i \log(y_i/x_i).$$

In information theory, the divergence of x from y is interpreted as a measure of how much statistical information (about some events) is imparted by a signal that shifts a person's beliefs about the likelihood of these events from an initial (prior) distribution, x, to a new (posterior) distribution, y. Appropriately, Proposition 0 states that a divergence of

---

[10]For a proof of the uniqueness of the logarithmic scale, see Savage (1971). When there are only two events, however, then certain other scales will work as well (Aczél and Pfanzagl, 1966). Toda (1963) suggested the quadratic, for example:

$$\text{Score} = (1-(1-x_i)^2); \quad i=1,2.$$

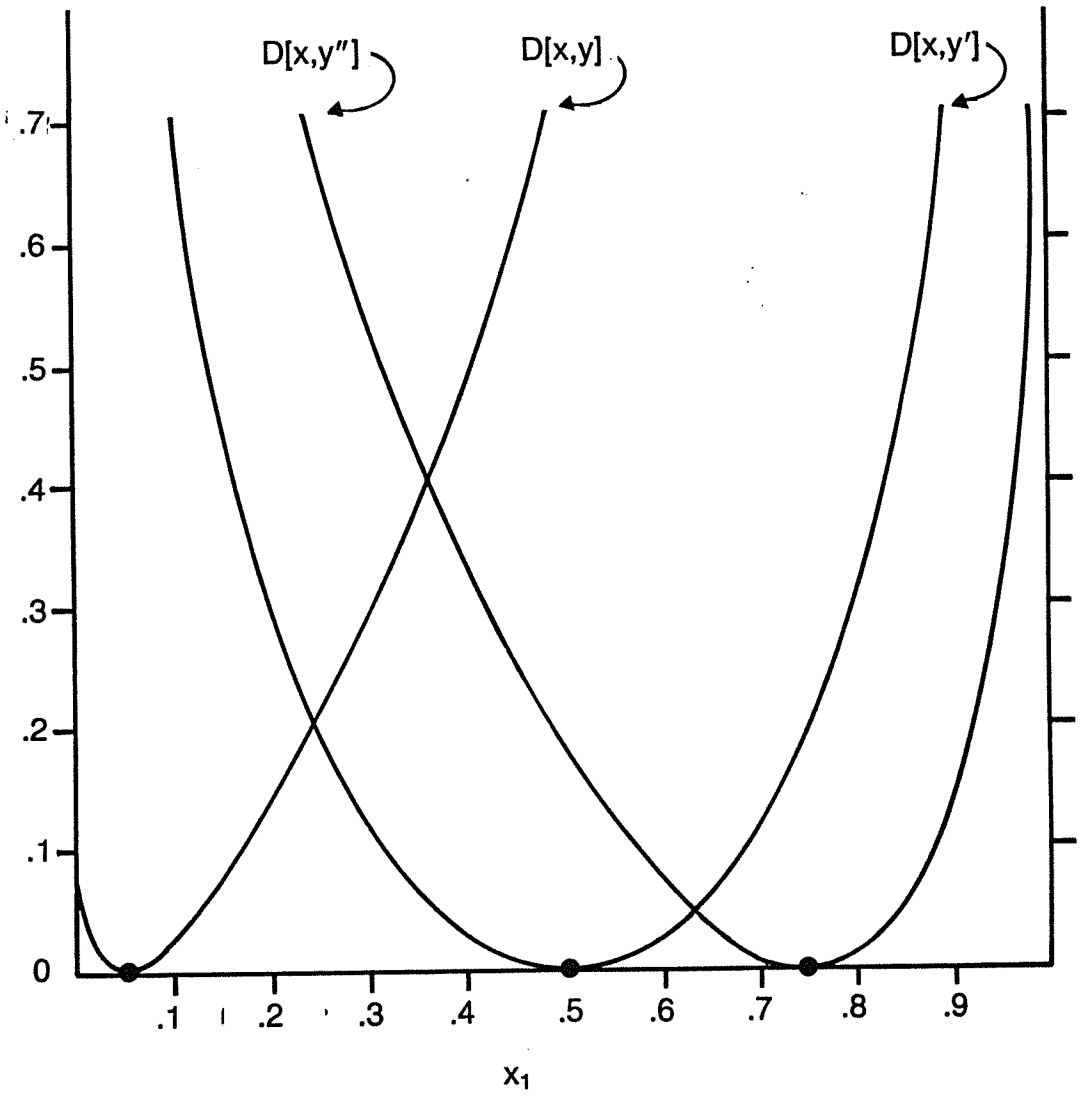


Figure 7: The divergence of the general binary distribution, $x = (x_1, 1-x_1)$ from three particular distributions:

$y = (.05,.95)$,
$y' = (.5,.5)$,
$y'' = (.75,.25)$.

one distribution from another is always greater or equal to zero, and only equals zero if the distributions coincide.[11] Figure 7 plots several divergences, for binary distributions.

Insert Figure 7 About Here

The equation for expected score

We now begin a systematic examination of this game, by stepping back and considering all possible strategies that the three players could adopt, and how these strategies would fare if played against each other. Formal ly, a strategy is a comprehensive rule indicating what a player will do, given the information at his disposal. Thus, a strategy for Player A, is a a list of conditional probability distributions, $r(\underline{a},\underline{A}|\alpha)$, one for each α, over possible vocabularies and descriptions:

$r(\underline{a},\underline{A}|\alpha)$ = Probability that A will present $\underline{A}$, and select $\underline{a}$ from $\underline{A}$ if he observes α.

For Player B, a strategy is a rule, b, that identifies an assessed probability distribution over elements in $\underline{A}$, as a function of observation β, and of course, $\underline{A}$:

$b(\underline{a}|\beta,\underline{A})$ = B's declared probability that A will select $\underline{a}$, if the presented vocabulary is $\underline{A}$ and if B observes β.

---

11 Note that this is not a symmetric relation, i.e., $D(x,y) \neq D(y,x)$. In statistics, $D(x,y)$ is somethimes referred to as the directed divergence, with the word divergence reserved for the (symmetric) sum $D(x,y) + D(y,x)$ (Kullback, 1954).

Likewise, a strategy for C is a rule, c, that also identifies an assessed distribution over $\underline{A}$, but with the omission of observation β:

$c(\underline{a}|A)$ = C's declared probability that A will select $\underline{a}$ if the presented vocabulary is $\underline{A}$.

A summary description of the game, and the players' strategies is shown in Figure 8.

Insert Figure 8 About Here

The strategies, r, b, c, are private, as are the observations, α and β. However, the expected consequences of particular strategies are public: because the probability distribution $p(\alpha,\beta)$ is common knowledge, each player can calculate the expected score when strategies, a, b, and c, are simultaneously played against each other:

$$EV(r,b,c) = \sum_{\alpha,\beta} p(\alpha,\beta) \sum_{\underline{a},\underline{A}} r(\underline{a},\underline{A}|\alpha)\log\left[\frac{b(\underline{a}|\beta,\underline{A})}{c(\underline{a}|\underline{A})}\right] .$$

The approach that game-theory takes in attempting to solve a particular game is to eliminate from consideration combinations of strategies that are not mutually credible in that some player would find it to his advantage to choose an action other than the one recommended by his strategy (Luce and Raiffa, 1957). The deletion of strategies may be conceptualized as a public process, since everyone can calculate whether a player will have reason to deviate. It is as if all possible combinations of strategies are initially listed on a public blackboard, from which individual combinations are crossed off, whenever at least one player would not wish to adhere to

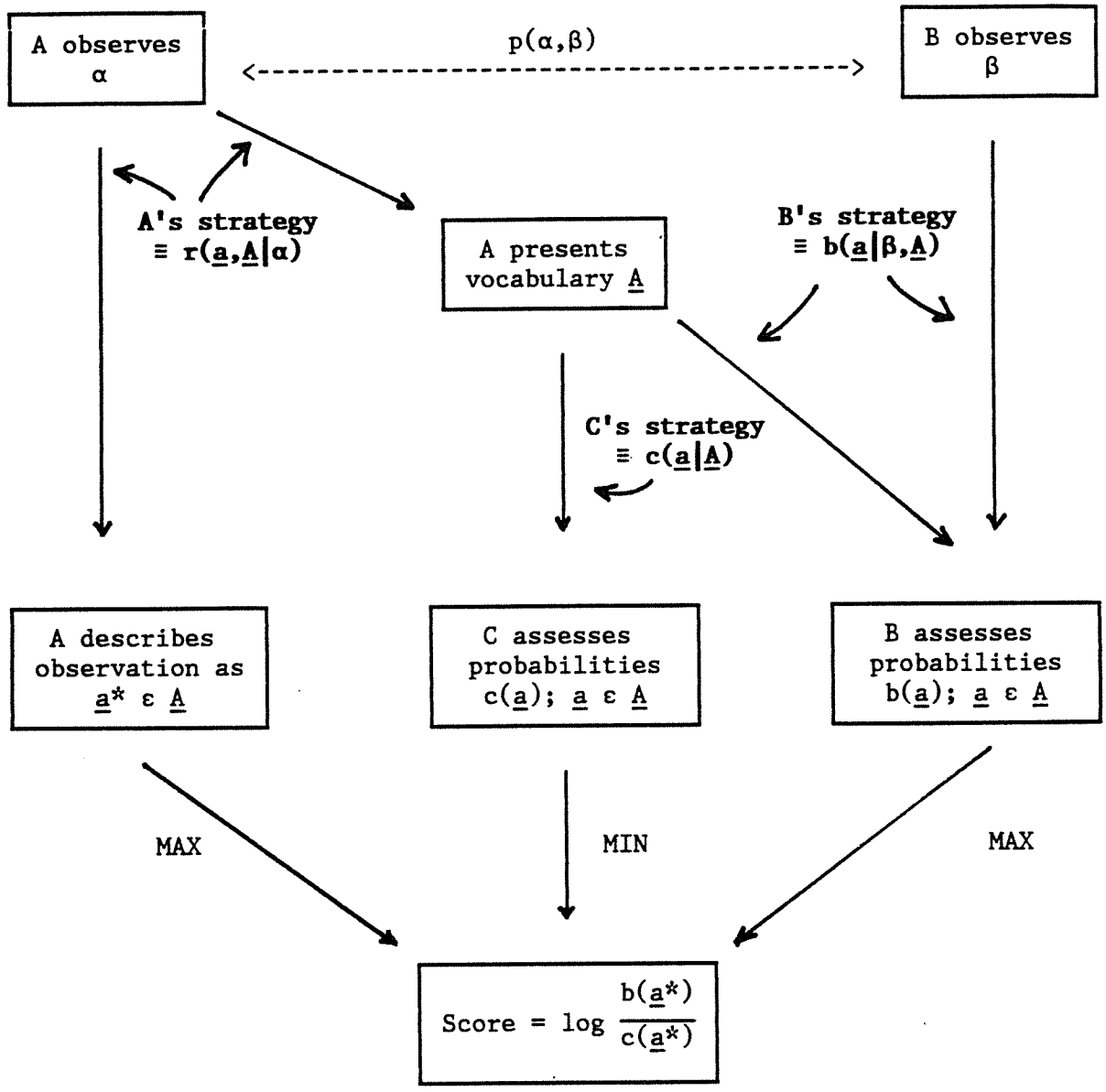


Figure 8: A schematic description of the communication game along with relevant notation.

his strategy. The surviving combinations of strategies -- those at which no player can gain by doing something else -- are then in a sort of (public) reflective equilibrium, and they are the candidates for solutions to the game.

We now bring this form of analysis to bear on our game. It will simplify the calculations if we extend A's strategy to a complete distribution over all four relevant variables $\underline{a}$, $\underline{A}$, α, β:

$$r(\alpha,\beta,\underline{a},\underline{A}) = \Pr\{\text{A selects } \underline{a} \text{ from } \underline{A},\ \text{A observes } \alpha,\ \text{B observes } \beta\}$$
$$= r(\underline{a},\underline{A}|\alpha)p(\alpha,\beta).$$

We will also write the associated marginal and conditional distribution of r in the usual way, e.g. $r(\underline{a})$, $r(\underline{a}|\beta)$, and so on. In this way, we redefine A's strategy as a selection of a four-variate distribution, $r(\alpha,\beta,\underline{a},\underline{A})$ from the set of all those whose marginal distribution on (α,β) concides with $p(\alpha,\beta)$,

$$r(\alpha,\beta) = p(\alpha,\beta),$$

and whose marginal distribution over $(\underline{a},\underline{A})$, conditional on α, is independent from β:

$$r(\underline{a},\underline{A}|\alpha) = r(\underline{a},\underline{A}|\alpha,\beta).$$

The starting point is the equation for expected score:

$$EV(r,b,c) = \sum_{\alpha,\beta,\underline{a},\underline{A}} r(\alpha,\beta,\underline{a},\underline{A}) \log\left[\frac{b(\underline{a}|\beta,\underline{A}}{c(\underline{a}|\underline{A})}\right].$$

We can rewrite the expression in the square brackets as the product of the five ratios:

$$\frac{p(\beta|\alpha)}{p(\beta)} \frac{r(\beta|\underline{a},\underline{A})}{p(\beta|\alpha)} \frac{p(\beta)}{r(\beta|\underline{A})} \frac{b(\underline{a}|\beta,\underline{A})}{r(\underline{a}|\beta,\underline{A})} \frac{r(\underline{a}|\underline{A})}{c(\underline{a}|\underline{A})},$$

(which, by applying Bayes' rule where appropriate, and canceling terms, indeed reduces to, $b(\underline{a}|\beta,\underline{A})/c(\underline{a}|\underline{A})$). The equation for expected score then separates into the corresponding five components,

$$EV(r,b,c) = \sum_{\alpha} p(\alpha) \sum_{\beta} p(\beta|\alpha) \log\left[\frac{p(\beta|\alpha)}{p(\beta)}\right] \qquad \text{(ED1)}$$

$$- \sum_{\alpha,\underline{a},\underline{A}} r(\alpha,\underline{a},\underline{A}) \sum_{\beta} r(\beta|\alpha) \log\left[\frac{p(\beta|\alpha)}{r(\beta|\underline{a},\underline{A})}\right] \qquad \text{(ED2)}$$

$$- \sum_{\underline{A}} r(\underline{A}) \sum_{\beta} r(\beta|\underline{A}) \log\left[\frac{r(\beta|\underline{A})}{p(\beta)}\right] \qquad \text{(ED3)}$$

$$- \sum_{\beta,\underline{A}} r(\beta,\underline{A}) \sum_{\underline{a}} r(\underline{a}|\beta,\underline{A}) \log\left[\frac{r(\underline{a}|\beta,\underline{A})}{b(\underline{a}|\beta,\underline{A})}\right] \qquad \text{(ED4)}$$

$$+ \sum_{\underline{A}} r(\underline{A}) \sum_{\underline{a}} r(\underline{a}|\underline{A}) \log\left[\frac{r(\underline{a}|\underline{A})}{c(\underline{a}|\underline{A})}\right] \qquad \text{(ED5).}$$

At the second summation, ED2, we made use of the fact that β and $(\underline{a},\underline{A})$ are independent, conditional on α: $r(\beta|\alpha,\underline{a},\underline{A})=r(\beta|\alpha)=p(\beta|\alpha)$; the other summations are straightforward. Given the way we have arranged terms, each of the five summations is, formally, an expected divergence of one distribution from another:

$$EV(r,b,c) = \sum_{\alpha} p(\alpha)\, D(p(\beta),p(\beta|\alpha)) \qquad \text{(ED1)}$$

$$- \sum_{\alpha,\underline{a},\underline{A}} r(\alpha,\underline{a},\underline{A})\ D(r(\beta|\underline{a},\underline{A}),p(\beta|\alpha)) \qquad \text{(ED2)}$$

$$- \sum_{\underline{A}} r(\underline{A})\ D(p(\beta),r(\beta|\underline{A})) \qquad \text{(ED3)}$$

$$- \sum_{\beta,\underline{A}} r(\beta,\underline{A})\ D(b(\underline{a}|\beta,\underline{A}),r(\underline{a}|\beta,\underline{A})) \qquad \text{(ED4)}$$

$$+ \sum_{\underline{A}} r(\underline{A})\ D[c(\underline{a}|\underline{A}),r(\underline{a}|\underline{A})]. \qquad \text{(ED5)}$$

By this rearrangement, we have resolved the expected score into a sum of five components, each of which must be either negative or positive, and, more importantly, each of which is controlled essentially by not more than one player.

As we shall see in the next section, the first term, which involves only the observational variables, α and β, is also the information-theoretic measure of mutual information between α and β; the remaining four terms are interpretable as specific types of error by one of the three players. In anticipation of the discussion that follows, Proposition 1 summarizes the expected score equation, and attaches labels to each of the five component parts:

Proposition 1 The expected score in the communication game equals:

(Mutual information between the observations of A and B) (ED1)

- (A's expected error in describing observations) (ED2)
- (A's expected error in constructing the vocabulary) (ED3)
- (B's expected assessment error) (ED4)
+ (C's expected assessment error) (ED5)

Derivation of optimal strategies

Let us start with Players B and C. The only component of expected score that B influences is ED4, a quantity that is no greater than zero. Since B wishes to maximize expected score, upon observing β and examining $\underline{A}$ he will report the distribution that minimizes $D(b(\underline{a}),r(\underline{a}|\beta,\underline{A}))$, and that, we know, is just the distribution $r(\underline{a}|\beta,\underline{A})$. Likewise, C, who wishes to reduce expected score, can only affect ED5, and after inspecting $\underline{A}$, he will report the distribution $c(\underline{a})$ that minimizes $D(c(\underline{a}),r(\underline{a}|\underline{A}))$, namely the distribution $r(\underline{a}|\underline{A})$. This confirms our expectation that the scoring rule will elicit a true assessment of B's and C's beliefs.

Proposition 2 The best response of B and C to the vocabulary presented by A, is to reveal the subjective probability distribution over descriptions that is consistent with the distribution, $r(\underline{a},\underline{A},\alpha,\beta)$, implied by that vocabulary:

$$b^{\circ}(\underline{a}|\beta,\underline{A}) = r(\underline{a}|\beta,\underline{A});$$
$$c^{\circ}(\underline{a}|\underline{A}) = r(\underline{a}|\underline{A}).$$

(The superscript "$\circ$" indicates that these are optimal strategies.)

This leaves us now with A. Although the distribution that he implicitly defines, r, enters into all five component parts of expected score, of these five he needs really to concern himself with only two. The first expected divergence, ED1, is entirely a function of the marginal distribution $r(\alpha,\beta)$, which is just the distribution over observations, $p(\alpha,\beta)$. It is therefore a constant, not affected by any player's actions. The last two terms, ED4 and ED5, are indeed functions of A's strategy, but if

players B and C do their part, these terms will vanish, irrespective of what strategy A adopts. The only two components of expected score that A has any effective control over, then, are ED2 and ED3, and these two terms determine how he should play the game.

The simpler component, ED3, is an expected divergence; because it is non-negative, and because it is subtracted from expected score, the best that A can hope to do is to reduce it to zero. To accomplish that, he must choose only those vocabularies for which the associated divergence, $D(p(\beta),r(\beta|\underline{A}))$, vanishes:

$$r^{\circ}(\underline{A}) > 0 \text{ implies: } D(p(\beta),r^{\circ}(\beta|\underline{A})) = 0,$$
$$\text{implies: } r^{\circ}(\beta|\underline{A}) = p(\beta).$$

The strategic principle that this expresses is that A must confine himself to vocabularies that reveal nothing essential about his observation, that is, nothing that might improve an outside observer's (really--C's) conjectures about B's observations.

The remaining term, ED2, has a complementary significance. Again, it is a sum of divergences, subtracted from expected score; to eliminate it A must describe an observation with only those elements for which the associated divergence equals zero:

$$r^{\circ}(\underline{a},\underline{A}|\alpha) > 0 \text{ implies: } D(r^{\circ}(\beta|\underline{a},\underline{A}),p(\beta|\alpha)) = 0$$
$$\text{implies: } r^{\circ}(\beta|\underline{a},\underline{A}) = p(\beta|\alpha).$$

Now the principle is exactly opposite to the one we just derived from ED3. In order to reduce ED2 to zero, A must describe an observation in such a way that the beliefs about B's observations that the description engenders duplicate the beliefs implied by that observation. In other words, upon receiving the description, an outside observer -- for example C -- should

know everything that A knows, about B's observation. This is in accord with our intuition that a perfect description of an event should be an informational substitute for the actual experience of that event.

Proposition 3 The objective of Player A is to present a vocabulary that completely conceals, and then a description that completely reveals the informational content of his or her observation.

There is a small family of rules for generating vocabularies and descriptions that is consistent with these twin requirements, and this family will be defined shortly; a most simple rule that will do as well as any other, however, is for Player A to present a complete vocabulary, assigning to each observation α, a unique identifying description, $\underline{a}^\circ(\alpha)$, which, A claims, will be chosen if and only if he observes α. In formal notation, such a strategy would be defined by:

$$r^\circ(\underline{a},\underline{A}|\alpha) = \begin{cases} 1 & \text{if } \underline{a} = \underline{a}^\circ(\alpha) \text{ and } \underline{A} = \underline{A}^\circ = \{\underline{a}^\circ(\alpha);\ \text{all } \alpha\}, \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, the first requirement, i.e., that the vocabulary offers no clues about α, is met, since the same comprehensive vocabulary will be presented irrespective of α; it is also true that $\underline{a}^\circ(\alpha)$ will pinpoint A's beliefs about β, since it is selected if and only if α occurs.

Therefore, if A states his intention to follow the fully-revealing strategy, r°, and backs it by presenting the comprehensive vocabulary, and if B and C accept this, and respond by revealing their probability distribution over elements of this vocabulary, then the last four components of expected score will disappear, leaving only the first component, which is the value of the game, EV°:

$$EV^\circ = EV(r^\circ, b^\circ, c^\circ) = \sum_{\alpha} p(\alpha) D(p(\beta), p(\beta|\alpha)). \qquad (1)$$

An example of optimal play

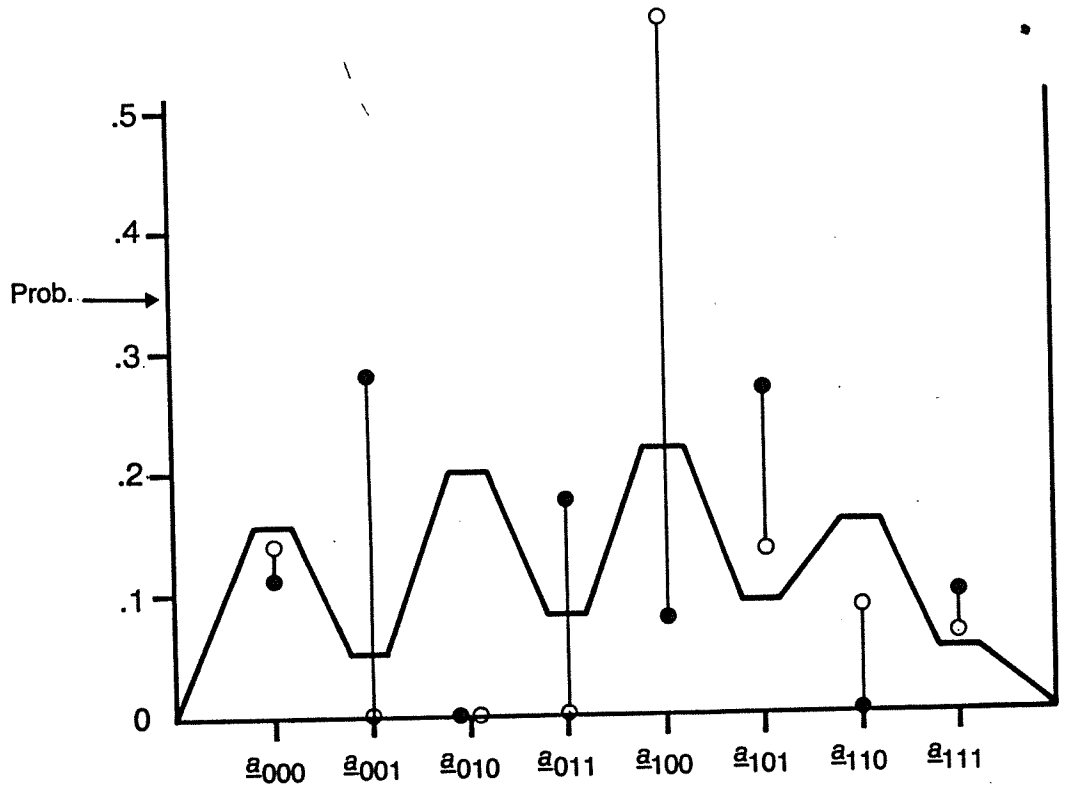
To make these conclusions more tangible, we will go back to our wine-tasting example, from Section 3, and see whether it is in fact optimal for Subject A to reveal all aspects of his observation. Suppose, then, that having observed α_{101}, Subject A defines a complete vocabulary, containing eight descriptions, $\underline{a}_{000}$ through $\underline{a}_{111}$, and that B and C have no reason to question his intention to describe observation α_{ijk} with its corresponding $\underline{a}_{ijk}$. The question now is: Can A improve his score by mislabelling α_{101} as something other than $\underline{a}_{101}$? Proposition 3 assures us that he cannot, which we will now check.

The relevant probability distributions (from A's perspective) are plotted in Figure 9. The solid line connects the probabilities, $p(\alpha_{ijk})$, that C will assign; the solid and open circles are the conditional probabilities derived from the two of B's observations that A considers possible. Specifically, the open circles indicate probabilities conditional on

Insert Figure 9 About Here

β_{100}, and the solid ones probabilities conditional on the (actual) β_{101}. Player A, then, will choose the description that maximizes the expected log ratio of B's to C's probabilities, which are given by the row of numbers below the x-axis. As we had hoped, the highest expected score (+1.06) accompanies the correct identification of α_{101} as $\underline{a}_{101}$.

At the end of Section 3, we noted that the description that A thinks will receive (on average) the greatest probability assessment by B is not $\underline{a}_{101}$ but $\underline{a}_{100}$. The reason why $\underline{a}_{100}$ is a poor choice for A in this game



Expected Score:	-.27	-∞	-∞	-∞	.11	1.06	-∞	.53

—— C's assessment, $c(\underline{a}_{ijk}) = p(\alpha_{ijk})$.

● B's actual assessment, $b(\underline{a}_{ijk}) = p(\alpha_{ijk}|\beta_{101})$.

○ B's other possible assessment, $b(\underline{a}_{ijk}) = p(\alpha_{ijk}|\beta_{100})$.

Figure 9: Player A's inferences about possible assessments of B and C, when he observes α_{101}. The expected scores, shown below the corresponding $\underline{a}_{ijk}$, are given by:

$$p(\beta_{100}|\alpha_{101})\log \frac{p(\alpha_{ijk}|\beta_{100})}{p(\alpha_{ijk})} + p(\beta_{101}|\alpha_{101})\log \frac{p(\alpha_{ijk}|\beta_{101})}{p(\alpha_{ijk})} .$$

The optimal label is $\underline{a}_{101}$. Infinite negative scores arise when there is a positive probability that B will assign zero-probability to a description.

(yielding only +.11), is that C also thinks it is fairly likely. Interesting, also, is that if A learned somehow that B's actual observation was β_{101}, then A's best description would no longer be $\underline{a}_{101}$ but instead $\underline{a}_{001}$ (yielding +2.41). But A doesn't know B's actual observation, and so must allow that B has observed β_{100}, and assigned zero-probability to $\underline{a}_{001}$.

5 EXPECTED SCORES AND INFORMATION THEORY

Readers familiar with information theory may have noticed that Equation 1 defines the mutual information between two random variables--in this case, the observational variables α and β. The statistics of information theory do crop up throughout the previous derivation, and we can use them to gain some additional insights into why the scoring system in this game requires A to completely describe his observation.[12]

The fundamental building block of information theory is the concept of the average uncertainty, or entropy, of a probability distribution. To use a concrete example, the uncertainty of A's observations is defined as:

$$U(\alpha) = \sum_{\alpha} p(\alpha) \log \frac{1}{p(\alpha)} , \qquad (2)$$

which is the expected "log-improbability" of the individual observations. If A's observations were equiprobable, then their uncertainty would equal the log of the number of observations.

---

[12] Garner (1962) is still the classic exposition of the relation of these ideas to psychological theory.

The main convenience that the logarithmic transform provides for this definition, is that it makes uncertainty additive over independent random events. For example, tossing a fair coin once creates 1 unit of uncertainty; tossing that same coin twice creates 2 units, and so on (it is customary to use the base-two logarithm). When the events are not independent, as is the case with the observations α and β, then the uncertainty of the joint distribution over pairs of events is somewhat less than the sum of the individual uncertainties,

$$U(\alpha,\beta) < U(\alpha) + U(\beta).$$

As we will see, the discrepancy between the joint uncertainty, $U(\alpha,\beta)$, and the sum of individual ones is one possible way of measuring the mutual information between two variables.

Let us now develop this concept of mutual information, in the context of our particular application. Consider B's uncertainty about A's observations. Prior to the observation period, B's beliefs about α are described by distribution $p(\alpha)$, and his uncertainty about α therefore is equal to the quantity defined in Equation 2. After observing β^*, B's beliefs change, from $p(\alpha)$ to $p(\alpha|\beta^*)$, and the new uncertainty is:

$$\sum_{\alpha} p(\alpha|\beta^*) \log \frac{1}{p(\alpha|\beta^*)} \quad . \tag{3}$$

The expected uncertainty that B has after the observation period is the average of Equation 3 over all observations, β, weighted by the probabilities of these observations.

$$U(\alpha|\beta) = \sum_{\beta} p(\beta) \sum_{\alpha} p(\alpha|\beta) \log \frac{1}{p(\alpha|\beta)} \quad . \tag{4}$$

The difference between Equations 4 and 2 is the average reduction in uncertainty from before and after the observation period, and it is the sensible definition of the amount of information about α that B can extract from his observation, on average. Remarkably, this quantity does not change if we reverse the roles of A and B, and write down the average reduction in uncertainty about β that A derives from his own observation, α:

$$U(\alpha) - U(\alpha|\beta) = \sum_{\alpha,\beta} p(\alpha,\beta) \log \frac{p(\alpha|\beta)}{p(\alpha)} \tag{5}$$

$$= \sum_{\alpha,\beta} p(\alpha,\beta) \log \frac{p(\alpha,\beta)}{p(\alpha),p(\beta)} \qquad \text{(by Bayes' rule)}$$

$$= U(\alpha) + U(\beta) - U(\alpha,\beta) \tag{6}$$

$$= \sum_{\alpha,\beta} p(\alpha,\beta) \log \frac{p(\beta|\alpha)}{p(\beta)} \qquad \text{(by Bayes' rule)}$$

$$= U(\beta) - U(\beta|\alpha). \tag{7}$$

Because of this symmetry, we refer to the uncertainty reduction, from the perspective of either party, as the mutual information, $I(\alpha;\beta]$, between α and β:

$$I(\alpha:\beta) = U(\alpha) - U(\alpha|\beta) = U(\beta) - U(\beta|\alpha).$$

It is convenient to represent these relations by means of set diagrams, such as the ones in Figure 10. In such a diagram, each variable being considered is assigned a set, whose area is drawn in proportion to the uncertainty of that variable. The union of two such sets corresponds to the joint uncertainty of the respective variables, while the

intersection corresponds to the mutual information between them. The two set differences, finally, correspond to the conditional uncertainties. It is easy to see from Figure 10 that the relations stated in Equations 5, 6, and 7 do hold.

Insert Figure 10 About Here

Let us now re-examine the equation for expected score. The last two terms, ED4 and ED5, are B's and C's expected assessment errors, and, since they have no direct information-theoretic interpretation, we will set them aside. The first term, ED1, is the mutual information between α and β, as can be seen by inserting the definition of a divergence, and comparing the result with Equation 5.

The remaining term, ED2, needs a little work. Since β conditional on α is independent from $\underline{a}$, we can replace $r(\beta|\alpha)$ with $r(\beta|\underline{a},\alpha)$ and rewrite the entire expression as:

$$\sum_{\underline{a}} r(\underline{a}) \sum_{\alpha} r(\alpha|\underline{a}) D(r(\beta|\underline{a}), r(\beta|\underline{a},\alpha)) \qquad \text{(ED2)}$$

This is a weighted sum, over descriptions, of terms,

$$\sum_{\alpha} r(\alpha,\beta|\underline{a}) \sum_{\beta} r(\beta|\alpha,\underline{a}) \log \frac{r(\beta|\alpha,\underline{a})}{r(\beta|\underline{a})} ,$$

each of which defins the mutual information between α and β, conditional on a specific description, $\underline{a}$. In information-theoretic terminology, the entire sum would be referred to as the mutual information of α and β, given $\underline{a}$, or, $I(\alpha:\beta|\underline{a})$.

The contributions of ED1, ED2, and ED3 to expected score can now be represented in the form of a set diagram, which is displayed in Figure 11.

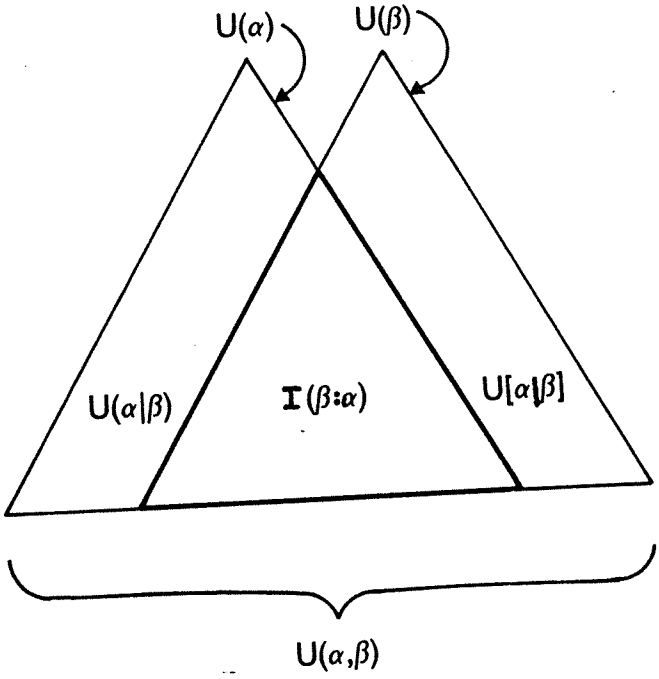


Figure 10: The set-theoretic representation of uncertainties.

Insert Figure 11 About Here

The figure contains four triangles, representing the uncertainty of the four relevant variables, α, β, $\underline{a}$, and $\underline{A}$. Although no pair of variables are necessarily independent, certain pairs are conditionally independent, given a third variable, and the arrangement of the triangles in the Figure reflects this. The vocabulary-triangle is contained within the description-triangle, because a description tags the vocabulary from which it originates (and the conditional uncertainty, $U(\underline{A}|\underline{a},\underline{A})$, is zero). More importantly, the description-triangle is placed is such a relation to the α- and β-triangles that the conditional independence of $\underline{a}$ and β, given α, is represented by the inclusion of the intersection of $\underline{a}$ and β within the intersection of α and β. The mutual information between $\underline{a}$ and β cannot, in other words, be greater than the mutual information between α and β. This makes intuitive sense. Clearly A's descriptions cannot provide more information about β than is already contained in the observations from which they are derived.

Let us now focus our attention on the triangle where α and β intersect; our previous analysis indicated that the area of this triangle -- the mutual information between α and β -- identifies the maximum expected score that A can sustain, if B and C assess probabilities without error. The shaded subset of this triangle is the expected score when A's play falls short of the ideal, but B and C are still assessing optimally. The difference between actual and potential expected score is the intersection of $\underline{A}$ and β, plus that portion of the α-β intersection that is not also contained in the $\underline{a}$-triangle. As these two areas correspond to ED3 and ED2, respectively, this is the decomposition that we have derived in the equation for expected score. Alternatively, we can observe that the shaded

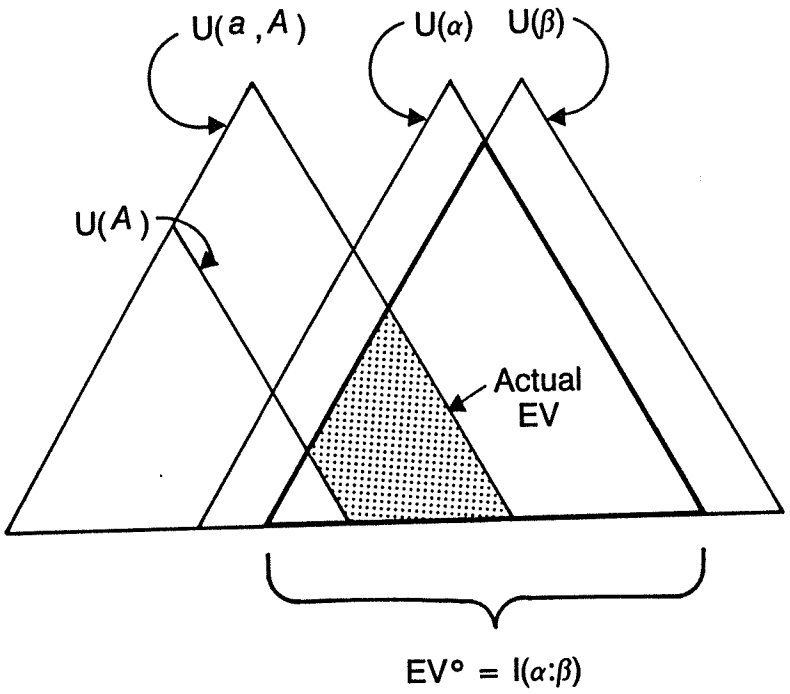


Figure 11: Decomposition of expected score, if B and C assess probabilities without error.

area is also the difference between the intersection of $\underline{a}$ and β, and the intersection of β and $\underline{A}$, and so express the actual expected score without referring to α:

$$EV(r,b^\circ,c^\circ) = I(\beta:\underline{a}) - I(\beta:\underline{A}).$$

In words: if Players B and C play without error, then the expected score for any strategy on the part of Player A equals the mutual information between descriptions and B's observations, minus the mutual information between vocabularies and observations.

We can now restate the conditions for optimal play by Player A. We already remarked that the mutual information between $\underline{a}$ and β cannot exceed the mutual information between α and β,

$$I(\beta:\underline{a}) \leq I(\beta:\alpha).$$

By definition, mutual information between vocabularies and observations cannot be negative,

$$I(\beta:\underline{A}) \geq 0.$$

The most that A can aim for, therefore, is to construct a vocabulary that signals nothing about β, but that is at the same time sufficiently flexible and comprehensive to express all aspects of the observation that might conceivably be of use in predicting β. If he succeeds in this, then no information is leaked through $\underline{A}$,

$$I(\beta:\underline{A}) = 0,$$

and no information is lost in passing from α to $\underline{a}$ so that the mutual information between $\underline{a}$ and β attains its theoretical maximum:

$$I(\beta:\underline{a}) = I(\beta:\alpha).$$

We now gather the results that have been derived so far.

Proposition 4 If A observes α^*, B observes β^*, and everyone plays without error, then:

(a) A's expectation of the score is the divergence of $p(\beta)$ from $p(\beta|\alpha^*)$.

(b) B's expectation of the score is the divergence of $p(\alpha)$ from $p(\alpha|\beta^*)$.

(c) C's expectation of the score is the mutual information between α and β.

(d) The actual score is:

$$\log\left[\frac{p(\alpha^*|\beta^*)}{p(\alpha^*)}\right];$$

(e) The values in (a), (b), and (c), and (d), are not altered when the roles of A and B are interchanged.

Let us first look at the expectation of A and B. The score that A expects is the divergence of his prior distribution over B's observation from the posterior distribution that he assesses after observing α^*; hence, it is a measure of how much information is imparted by α^*. If the prior and posterior distributions coincide, then α^* carries no information, and the expected score will equal zero. The same interpretation applies to B's expectation, as given in (b).

The actual score is defined in (d) as the log ratio of the conditional and marginal probabilities of α^*, that is, the log ratio of B's and C's

beliefs. It is instructive to rewrite this ratio in a way that expresses the symmetry between the roles of A and B:[13]

$$\frac{p(\alpha|\beta)}{p(\alpha)} = \frac{p(\alpha|\beta)\ p(\beta|\alpha)}{p(\alpha,\beta)}$$

The actual score, which we will label $v^{\circ}(\alpha,\beta)$, measures the accuracy of A's and B's mutual knowledge of each other's observations, namely $p(\beta|\alpha)$ times $p(\alpha|\beta)$, against C's knowledge, $p(\alpha,\beta)$:

$$v^{\circ}(\alpha,\beta) = \log\left[\frac{p(\alpha|\beta)\ p(\beta|\alpha)}{p(\alpha,\beta)}\right] .$$

The scores that A, B, and C expect, are the expectations of $v^{\circ}(\alpha,\beta)$ over β, α, and (α,β). Although these expectations are all greater than or equal to zero, the actual score is sometimes negative, as the example below illustrates:

	β_1	β_2	
α_1	.30	.00	.30
α_2	.30	.10	.40
α_3	.15	.15	.30
	.75	.25	

$p(\alpha,\beta)$

	β_1	β_2	
α_1	.42	--	.42
α_2	.00	.00	.00
α_3	-.58	1.00	.21
	.05	.60	**.19**

$v^{\circ}(\alpha,\beta)$

The marginal values in the second matrix give the expected scores conditional on individual observations, α, or β (i.e., the divergences mentioned

---

13 This is a restatement of Bayes' rule: $p(\beta|\alpha) = p(\alpha,\beta)/p(\alpha)$.

in Proposition 4); the value in the box is the unconditional expected score (or C's expectation). Notice that observation α_3 is informative, even though it yields even odds for β_1 vs. β_2, while observation α_2, which points strongly to β_1, has no value. This is because α_2 does not change the prior (marginal) odds, which are 3:1, while α_3 shifts them considerably.

Mutually uninformative distinctions and aspects

In the discussion following Proposition 3, we established that A can maximize expected score by creating a descriptive system that allows him to uniquely identify each possible observation. The question we now address is whether A can in some circumstances afford to be less than fully comprehensive. There are, it turns out, two cases in which A can give incomplete information about his observation without suffering any loss in expected score.

Case I: If two observations, α and α', induce the same beliefs about B's observations,

$$p(\beta|\alpha) = p(\beta|\alpha'), \quad \text{for all } \beta,$$

then A may cover them with the same description,

$$\underline{a}^\circ(\alpha) = \underline{a}^\circ(\alpha'),$$

with no loss in score.

This just restates our earlier conclusion, that observations with the same belief-signature need not be differentiated in games of pure strategy.

Case II If an aspect, (α), of the actual observation (i.e. set of observations containing the actual one) carries no information about B's observations,

$$p(\beta|\alpha \in (\alpha)) = p(\beta),$$

then Player A can omit that aspect from his vocabulary set without loss in expected score.

What this means is that A can make aspect (α) public, by prefacing his vocabulary with an introductory statement, and then restrict his vocabulary set to descriptions of observations that are characterized by (α). The next example illustrates this.

	β_1	β_2	
α_1	.2	.0	.2
α_2	.2	.1	.3
α_2	.1	.2	.3
α_4	.0	.2	.2
	.5	.5	

$p(\alpha,\beta)$

	β_1	β_2	
α_1	1.0	--	1.00
α_2	.42	-.58	.09
α_3	-.58	.42	.09
α_4	--	1.0	1.00
	.45	.45	.45

$v^\circ(\alpha,\beta)$

Suppose that A observes α_1; a possible way to play the game would be to write the full vocabulary set, corresponding to the four observations, yielding an expected score of 1.00. He can, however, save some effort, in the following way. The four possible observations are characterized by two complementary aspects,

$$(\alpha)_H = \{\alpha_1,\alpha_4\},$$
$$(\alpha)_L = \{\alpha_2,\alpha_3\},$$

which we can call the High-confidence, and Low-confidence aspects. These aspects do not say anything about B's observation since knowing that an observation is High or Low confidence does not alter the prior 50-50 distribution over B's observations. A can now announce that his observation is High-confidence, define two descriptions, corresponding to α_1 and α_4, and play the resulting game, whose expected score is still 1.0:

	β_1	β_2	
α_1	.5	.0	.5
α_4	.0	.5	.5
	.5	.5	

$p(\alpha,\beta|\alpha \in (\alpha)_H)$

	β_1	β_2	
α_1	1.0	-	1.0
α_4	-	1.0	1.0
	1.0	1.0	**1.0**

$v^\circ(\alpha,\beta|\alpha \in (\alpha)_H)$

If two aspects are uninformative, then so are the aspects defined by taking the union, or intersection of observations that they characterize. For this reason one can always find, for each observation, a unique (possibly empty) string of aspects that can be made public without loss in score.

The two cases discussed here are the only qualifications on the claim that A must fully reveal his observation. Case I states which distinctions may remain private, and Case II which may be made public prior to actual play. It should be repeated, however, that A cannot lose by exhaustively listing all observations, and should he be in any doubt about the applicability of the conditions in I and II to a particular instance, it would be safer for him to include the marginally informative distinctions in the vocabulary.

6 THREE FORMS OF INTROSPECTING ERROR

In this section, we classify and, interpret several of the imperfections commonly attributed to introspective judgment as specific errors on the part of Player A. We assume throughout, however, that B and C are aware of these imperfections, and adjust their probability assessments accordingly.

Inconsistency/Ambiguity

The most widespread concern about introspection is that it is difficult for a person to consistently describe phenomena that are directly accessible to only that person. How does one know whether the same phenomenal datum is not being described differently on different occasions in spite of the best effort by the introspecting individual?

Within the framework of the introspecting game, an error of this type would be represented by the presence of randomization in A's strategy. A concrete example may help illustrate this. Suppose that it is known that both A and B will observe one of two equiprobable things, and that what they observe is common knowledge between them:

	β_1	β_2	
α_1	.5	.0	.5
α_2	.0	.5	.5
	.5	.5	

$p(\alpha,\beta)$

	β_1	β_2	
α_1	1.0	-	1.0
α_2	-	1.0	1.0
	1.0	1.0	1.0

$v^\circ(\alpha,\beta)$

The introspecting player, A, accordingly defines two descriptions, $\underline{a}_1$ and $\underline{a}_2$, but fails to use them consistently. To be specific, let us say he

mislabels α_1 as $\underline{a}_2$ 10% of the time, and mislabels α_2 as $\underline{a}_1$ 20% of the time. In matrix notation, his strategy would be written as:

$$\begin{array}{c} \\ \alpha_1 \\ \alpha_2 \end{array} \begin{array}{c} \begin{array}{cc} \underline{a}_1 & \underline{a}_2 \end{array} \\ \begin{bmatrix} .9 & .1 \\ .2 & .8 \end{bmatrix} \end{array}$$

$r(\underline{a}|\alpha)$

The joint distribution of descriptions and B's observations, which determines expected score, is then:[14]

$$\begin{array}{c} \\ \underline{a}_1 \\ \underline{a}_2 \\ \\ \end{array} \begin{array}{c} \begin{array}{cc} \beta_1 & \beta_2 \end{array} \\ \begin{bmatrix} .45 & .10 \\ .05 & .40 \end{bmatrix} \\ \begin{array}{cc} .50 & .50 \end{array} \end{array} \begin{array}{c} \\ .55 \\ .45 \\ \\ \end{array}$$

$r(\underline{a},\beta)$

$$\begin{array}{c} \\ \underline{a}_1 \\ \underline{a}_2 \\ \\ \end{array} \begin{array}{c} \begin{array}{cc} \beta_1 & \beta_2 \end{array} \\ \begin{bmatrix} .71 & -1.46 \\ -2.17 & .83 \end{bmatrix} \\ \begin{array}{cc} .42 & .37 \end{array} \end{array} \begin{array}{c} \\ .32 \\ .50 \\ \boxed{.40} \end{array}$$

$v^{\circ}(\underline{a},\beta)$

The matrix on the right gives the scores, for particular combinations of $\underline{a}$ and β, if Players B and C assess without error. As we can see, a moderate probability of mislabelling diminishes the expected score by more than 60%. The sensitivity of the scoring system to inconsistency is due to the large losses (i.e., the off-diagonal entries in the matrix) that A and B suffer on those occasions when an inappropriate description is selected by A.

---

[14] A direct way to calculate this matrix in more complicated examples is to multiply the transpose of the strategy matrix by the $p(\alpha,\beta)$ matrix:

$$r(\underline{a},\beta) = r(\underline{a}|\alpha)^{t} p(\alpha,\beta).$$

Generally, the rarer these errors, the greater the cost in making them, as B will be caught with a smaller assessed probability of them occurring.

In these last few remarks, I have discussed inconsistent use of descriptions as a phenomenon that might be exhibited through repeated play of the introspecting game. The conceptual analogue of this type of error in the single-play game would be the presentation, by A, of a vocabulary that creates some confusion (in the mind of B and C) about the process by which descriptions are selected. What matters, in both cases, is the impression of randomness; whether this perception arises out of inconsistenties observed in repeated play, or whether it arises at the moment when A presents an ambiguous or poorly understood vocabulary, is not itself important.

The top left panel in Figure 12 diagrams the loss in score due to inconsistency/ambiguity, using the set representation of uncertainties. The description triangle is shifted away from the observation triangles, leaving a gap between potential and actual expected score.

Insert Figure 12 About Here

Incompleteness

The top right panel in Figure 12 captures a conceptually distinct form of introspecting error, which occurs when the introspecting subject describes consistently, but omits to include in his description some relevent aspects of phenomenal experience. The inclusion of the description-set within the observation-set indicates that the uncertainty of descriptions given observations is zero, or that in this system every observation is described in exactly one way. However, some observations having different implications for B's observation must receive identical

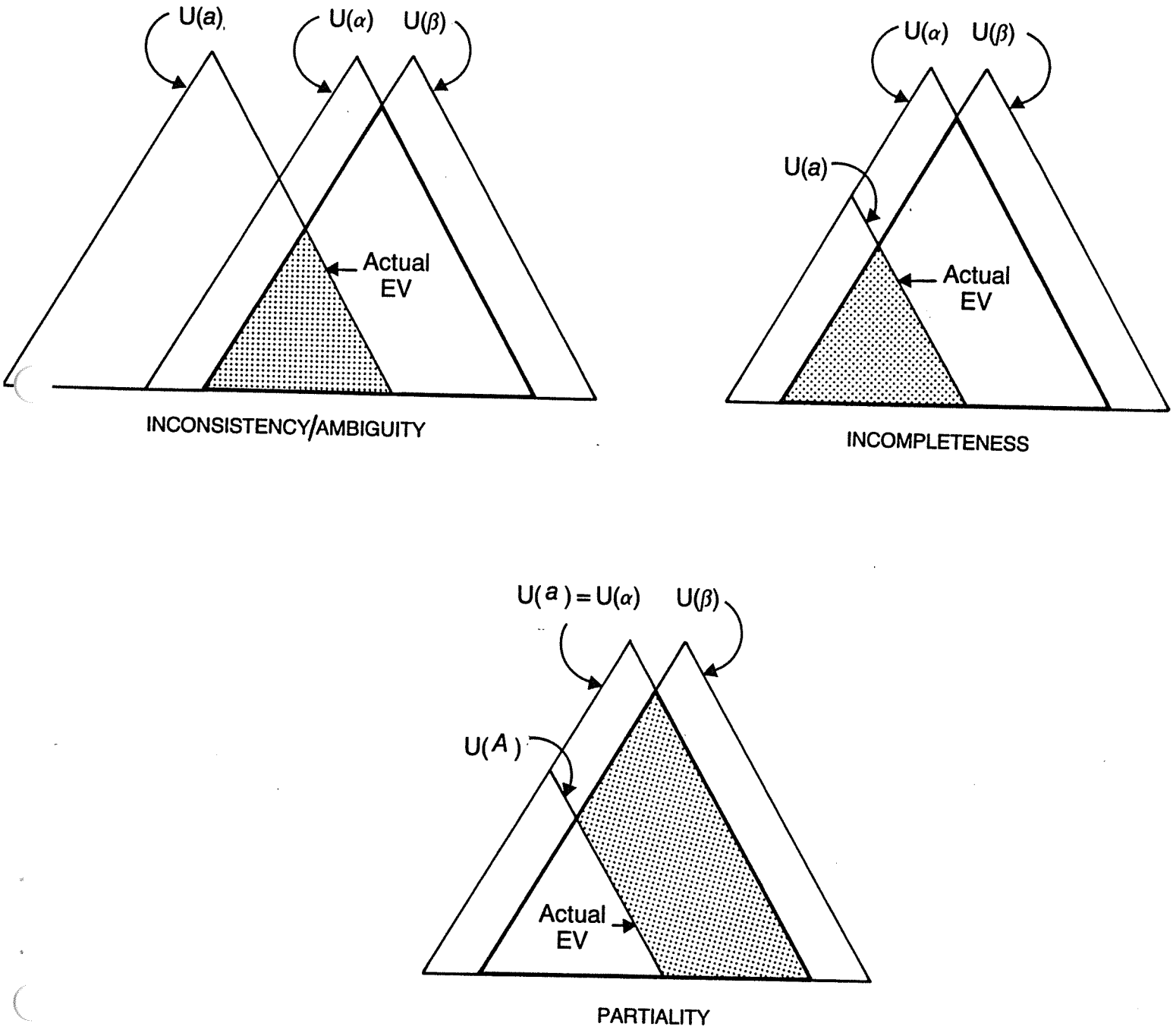


Figure 12: Three types of introspecting errors.

descriptions, or otherwise the gap between actual and potential expected score would disappear.

The example below illustrates how omission of relevant information is penalized by the scoring system. In the example, it is known in advance that Player B will observe one of three equiprobable things: β_1, β_2, or β_3. Player A's observations, however, are more differentiated; there are six distinct ones, and each one specifies a two-to-one-to-zero odds, in some particular order, over the three observations of B. We can think of A's observations as singling out one of B's observations as the most likely one, and then excluding one of the remaining two as impossible.

	β_1	β_2	β_3	
α_1	2/18	1/18	0	1/6
α_2	2/18	0	1/18	1/6
α_3	1/18	2/18	0	1/6
α_4	0	2/18	1/18	1/6
α_5	1/18	0	2/18	1/6
α_6	0	1/18	2/18	1/6
	1/3	1/3	1/3	

$p(\alpha,\beta)$

	β_1	β_2	β_3	
α_1	1.0	.0	-	.67
α_2	1.0	-	.0	.67
α_3	.0	1.0	-	.67
α_4	-	1.0	.0	.67
α_5	.0	-	1.0	.67
α_6	-	.0	1.0	.67
	.67	.67	.67	**.67**

$v^\circ(\alpha,\beta)$

Rather than present a full vocabulary, consisting of descriptions for all six observations, A decides to simplify, and provide only three possible descriptions, identifying the observation of B that is judged to be most likely. As a result α_1 and α_2 fall under the same description ($\underline{a}_1$), as do α_3 and α_4, and α_5 and α_6.

	β_1	β_2	β_3	
$\underline{a}_1$	4/18	1/18	1/18	1/3
$\underline{a}_2$	1/18	4/18	1/18	1/3
$\underline{a}_3$	1/18	1/18	4/18	1/3
	1/3	1/3	1/3	

$r(\underline{a},\beta)$

	β_1	β_2	β_3	
$\underline{a}_1$	1.0	-1.0	-1.0	.33
$\underline{a}_2$	-1.0	1.0	-1.0	.33
$\underline{a}_3$	-1.0	-1.0	1.0	.33
	.33	.33	.33	**.33**

$v^{\circ}(\underline{a},\beta)$

By bringing his vocabulary in line with B's observations, and eliminating the residual information concerning the two less likely observations, A has cut the expected score in half. With a full vocabulary, B's guess is always at least as good as that of C; with the simplified vocabulary, there is a one-third chance that the $(\underline{a},\beta)$ pair will fall of the diagonal, in which case C's guess of 1/3 is better than B's assessment, 1/6.

The general point that is being made here is that attempts to trim the vocabulary so as to bring it into a 1:1 alignment with the other person's perceptions will reduce expected score, unless that alignment is already present in the distribution $p(\alpha,\beta)$.

Partiality

The third type of introspecting error arises when the vocabulary that A presents cannot be used to describe certain observations other than the actual one, and so provides an impoverished or distorted context for interpreting the description of the actual observation. In calling such a vocabulary partial, I make use of both senses of the word: the vocabulary could be partial in that it makes no provision for describing some observations that might have, but in fact did not occur, or it could be partial to some observations over others by making descriptions of the favored observations in some way easier or more natural. As a result, when such a

vocabulary is presented by A, then both B and C may be able to exclude or assign very low probability to some of A's observations. The scoring rule in the introspecting game ensures that any such implicit leaking of information always helps Player C more than it does B.

We turn again to a numerical example for illustration. A and B are in almost perfect agreement about two possible observations:

$$p(\alpha,\beta)$$

	β_1	β_2	
α_1	.45	.05	.50
α_2	.05	.45	.50
	.50	.50	

$$v^{\circ}(\alpha,\beta)$$

	β_1	β_2	
α_1	.85	-2.32	.53
α_2	-2.32	.85	.53
	.53	.53	$\boxed{.53}$

A's strategy, now is to describe the two observations consistently, as $\underline{a}_1$ and $\underline{a}_2$; however, he presents these observations in a way that suggests strongly what he has in fact observed. This suggestion could be created by the order in which the two descriptions are presented, by the attention with which they are constructed, or by some explicit hints. If we label the vocabulary partial to α_1 as $\underline{A}_1$, and the one partial to α_2 as $\underline{A}_2$, then the strategy that A uses can be specified by a matrix such as the one below:

$$r(\underline{a},\underline{A}|\alpha)$$

	$\underline{a}_1,\underline{A}_1$	$\underline{a}_2,\underline{A}_1$	$\underline{a}_1,\underline{A}_2$	$\underline{a}_2,\underline{A}_2$
α_1	.8	.0	.2	.0
α_2	.0	.2	.0	.8

This matrix gives 20% chance that A will present a misleading vocabulary. As a result, B and C can use the information provided by the vocabulary to

update their beliefs about A's observation; C, who starts with even odds, updates to 4:1, while B, who starts with 9:1 odds, updates to either 9:4 or 36:1, depending on whether the vocabulary confirms or contradicts the observational information. The full implications of this joint reassessment by B and C can be calculated from the joint distribution of descriptions/vocabularies and B's observations:

	β_1	β_2	
$\underline{a}_1,\underline{A}_1$	.36	.04	.40
$\underline{a}_2,\underline{A}_1$	.01	.09	.10
$\underline{a}_1,\underline{A}_2$	.09	.01	.10
$\underline{a}_2,\underline{A}_2$	.04	.36	.40
	.50	.50	

$r(\underline{a},\underline{A},\beta)$

If, for example, the presented vocabulary is $\underline{A}_1$, then only the top half of this matrix is relevant, giving us a game with the following structure:

	β_1	β_2	
$\underline{a}_1$	.72	.08	.80
$\underline{a}_2$	.02	.18	.20
	.74	.26	

$r(\underline{a},\beta|\underline{A}_1)$

	β_1	β_2	
$\underline{a}_1$	.28	-1.38	.11
$\underline{a}_2$	-2.89	1.79	1.32
	.19	.81	**.35**

$v^{\circ}(\underline{a},\beta|\underline{A}_1)$

The analysis is the same for the other case, when the presented vocabulary is $\underline{A}_2$, so that we arrive at .35 as the expected score for the entire strategy, which is appreciably less than the potential expected score, .53.

A closer examination of the scores in the matrix reveals a curious fact, namely, that A's expected score when he presents a misleading vocabulary, such as $\underline{A}_1$ for observation α_2, is 1.32, which is much better than the .53 that can be obtained with an impartial vocabulary. What prevents A then from improving his score by dropping misleading hints about his observations? In a repeated game, this trick clearly won't work, because the presented vocabulary will only cause B and C to alter their beliefs if the choice of $\underline{A}_1$ or $\underline{A}_2$ has proven to be a reliable predictor of selected descriptions. If A always chooses the misleading vocabulary, then that vocabulary will no longer mislead, but will instead correctly signal A's subsequent description.

The argument in the one-shot game is somewhat more delicate. Now B and C have no past history of play to go by, but must instead reason through A's motives in providing clues about his selection. In doing so, they will see that depending on whether the clues are misleading or not, A's expected score is either 1.32 or .11. Since giving no clues will yield an expectation of .53 in a one-shot game, A has absolutely no reason to push the beliefs of B and C in favor of the description he will choose. But then his suggestion that he will choose $\underline{a}_2$, say, will lack all credibility, since he will only gain by this if he plans to choose $\underline{a}_1$. Upon reflection, then, A is seen to lack the ability to offer credible hints about his description, as it would only be to his advantage to make such statements if he intends to go against them.

Therefore, if we adhere strictly to our assumption that A, B, and C are fully rational and mutually aware of this fact, then in a one-shot game it is not really possible for A to present a partial vocabulary, as such a presentation could not made credible. Of course this is not true in a repeated game, as A could - intentionally or not - signal his descriptions

in many ways, and would diminish the long run score as a result. The loss in score from this type of error is represented in the bottom panel of Figure 12, where the potential expected score has a portion take out, corresponding to the mutual information between $\underline{A}$ and β, which is the signalling power of the vocabularies that A chooses.

Summary

In the previous discussion, we have treated the three types of introspective error as deliberate stratagems by Player A, and have shown that when B and C optimally adjust their expectations about A's actions, those stratagems will cause a reduction in expected score. Demonstrating this involved us in some fairly subtle game-theoretic arguments about credibility, etc., and may have obsured the essential strategic simplicity of the game we have been analyzing. Player A must present a description of what he has observed, and supply a context for interpreting the description, in the form of a system for describing the entire range of observational possibilities. B and C, in turn, supply probability distributions over possible descriptions, and are scored according the accuracy of their predictions. In playing the game, the players need not duplicate for themselves the mathematical reasoning the guides rational play. Rather, the scoring rule gives a critical piece of qualitative information, which is that there is no reason for A to be less than fully complete and consistent in his descriptions, and no reason for B and C to present distorted probabilistic assessments. The immunity of the scoring rule from strategic manipulation and misrepresentation can just as well be explained to the players before play begins; in that way, they are left free to concentrate on the formulation of correct descriptions and assessments.

7 ELICITATION-BY-ASPECTS AND OTHER VARIANTS

Sequential procedures

As it stands, the game we have developed has a serious drawback for practical application, in that for even a moderately complex structure of mutual knowledge it requires elicitation of subjective probability distributions over a prohibitively large number of events. For example, if Player A generates his descriptions by means of twenty independent binary aspects, each of which could either apply or not apply to his observation, then the set of possible descriptions will consist of over a million items (i.e., 2^{20}). There is no problem in recording the correct description, since it is just a sequence of twenty True/False values, but how is one to extract B's and C's subjective probability distributions over all possible descriptions? In this section we examine a procedure that resolves this difficulty by eliciting the original description, and the corresponding probabilities, in a series of smaller, more manageable steps.

This procedure, which we call the sequential introspecting game, differs from the standard one in two respects. First, instead of defining his vocabulary as an unstructured list of descriptions, A is encouraged to organize his knowledge in the form of a tree structure, whose endpoints correspond to possible completed descriptions of his observations. In this way, by selecting a description, A also selects a path leading from the top node to the endpoint corresponding to that description. We can imagine the paths leading from the top node as performing a preliminary, general classification of the object, while those leading from lower nodes as providing more and more detailed information about it.

Second, the subjective probability distributions of B and C are not elicited all at once, but in the following sequential manner: Starting

with the top node, B and C announce a probability distribution over all paths leading from that node; having done that, the correct path from that node is revealed, and the score computed using the same formula as in the original game.[15] This brings them down to the next node, where they play the same game over again. The final score for the game is the cumulated total of scores for all the node-games played on the way down to the endpoint.

Definition 4 An N-stage communication game consists of N rounds, each of which is played and scored by the same rules as apply to the single-stage game. The description selected by A at stage k is revealed before round k+1 commences, but the cumulated score is only made public at the end.

This is the general idea for the sequential game. I will, in fact, examine a more specialized version of it, in which each node is identified with an aspect, which could apply or not apply, thus yielding two possible descending paths. But before doing that, I want to state an important general property that sequential games have: If play is optimal, then the expected score in the game is determined solely by the subjective probability distributions over the endpoints, that is, the original descriptions of Player A; any tree-structure that generates the same distributions over these endpoints should yield the same expected score.

---

15 Why is the correct answer revealed? If it were not, then A and B could cumulate the score indefinitely by repeating the same vocabulary at each node.

Proposition 5 If play is without error at each stage, then the score in a sequential game does not depend on the number of stages allowed, nor on the order in which descriptive information is revealed by Player A.

The proposition is proved in Appendix I.[16]

Elicitation-by-aspects

A simple version of the sequential game is one in which A is instructed to present his description in the form of a sequence of statements that may be either true or false. After A has privately assigned a truth value to each statement, then, beginning with the first statement in the sequence, B and C are asked assess their probability that X has labelled the statement as true. The scoring is done in the usualy way, by adding (for B) and subtracting (for C) logarithms of the probabilities reported for all correct truth-values.

As the statements, and the corresponding truth-values are revealed, one-by-one, an increasingly detailed picture of A's observation is placed on the table, so to speak, for all to see. The game ends when A cannot

---

16 The mathematical support for this result derives from the conditional additivity of the mutual information statistic. Specifically, if $\underline{x}$,$\underline{y}$, and $\underline{z}$ are three random variables, then the mutual information between variable $\underline{z}$, and the product variable $(\underline{x},\underline{y})$, can be split into two components, the mutual information between $\underline{z}$ and $\underline{x}$, and the mutual information between $\underline{z}$ and $\underline{y}$, given $\underline{x}$:

$$I((\underline{x},\underline{y});\underline{z}) = I(\underline{x}:\underline{z}) + I(\underline{y}:\underline{z}|\underline{x}).$$

We can think of the left side of this equation as the potential expected score in the single stage game, where A's description is $(\underline{x},\underline{y})$ and B's observation is $\underline{z}$. The two terms on the right are then the expected score where A offers a partial description $\underline{x}$ in the first-stage of a two-stage game, and subsequently completes his description with $\underline{y}$, in the second stage.

think of any further statement whose truth or falseness is not implied by the statements already made public. At that point the description is complete, because C, who did not participate in the observation period, knows as much about what transpired there as does B, who was present.

Notice that not only B and C, but A, too, has less work to do in the sequential game, because he no longer needs to specify the vocabularies for nodes that he will not reach. For example, if the observation is a wine-tasting (and that is all that C knows), and if the wine happens to be red, then A does not have to provide his classification of attributes for white wines, as he would have to do in the one-shot game. He can place the red/white alternative at the top node, and then, having encoded the attribute "red," forget about white wines altogether.

All this convenience has a cost, namely, that the investigator will not obtain information about possible descriptions, and corresponding beliefs, for nodes that are off the path selected by A (e.g., about white wines, in the above example). In designing a sequential game, therefore, the investigator has to compromise between the simplicity of the game, and the amount of information he wants to receive about observations other than the actual one.

Limited-vocabulary games

When the number of stages is not limited, then the elicitation-by-aspects procedure we just described will yield the same score as the single-stage game. If, however, there is a limit to the number of aspects that A can produce, then the maximum score that A and B will attain may underestimate the mutual information between their observations. The problem is not one specific to sequential games, but would

arise in the same form in the one-stage game, if we placed a limit on the total number of descriptions contained in the vocabulary.

The only player affected by a restriction in vocabulary is A, who must find which partition of his observations into the requisite number of descriptions is most informative. Formally, he must consider vocabularies of size no greater than some integer N, and select the one that maximizes the mutual information between descriptions in that vocabulary and B's observations. This has somewhat the flavor of a categorization exercise, since A has to decide, essentially, which subsets of his observations would best be covered by a common description.

The example below illustrates the problem. A has three possible observations, as does B, but the number of descriptions is limited to two.

	β_1	β_2	β_3	
α_1	.2	.1	.0	.3
α_2	.1	.2	.0	.3
α_3	.0	.2	.2	.4
	.3	.5	.2	

$p(\alpha,\beta)$

	β_1	β_2	β_3	
α_1	1.15	-.58	-	.57
α_2	.15	.42	-	.33
α_3	-	.00	1.32	.66
	.82	.05	1.32	**.54**

$v^{\circ}(\alpha,\beta)$

A has to choose, then, which pair of observations will be described in the same way. Without calculating, it appears that α_1 and α_2 have the most similar informational content, since they both exclude β_3 from consideration. Indeed, if α_1 and α_2 are assigned the same description, the expected score drops only to .49. In contrast, combining α_2 with α_3 yields an expected score of .24, and the most dissimilar pairing, α_1 with α_3, an expected score of .12. In any experimental implementation, the number of aspects allowed A will necessarily be limited, which means that the final

score for the game will be somewhat underestimate the true mutual information between observations of A and B. For now, we just make note of this problem, and postpone a fuller discussion for another article.

A simulation of play

We will now illustrate Proposition 5 with an actual calculation, using, again, the wine-tasting example (Figure 6). In the top panels of Figure 13 we see the initial positions of B and C: B can restrict wines to the smaller shaded region (which corresponds to his actual observation β_{101}), while C has no information, yet. Player A then presents his first

Insert Figure 13 About Here

aspect $(\alpha)_1$, and privately gives it the (correct) Yes-designation. The first pair of assessed probabilities by B and C are .44 and .52, respectively, which gives C a slight advantage at this point.

$$\text{Score after first aspect} = \log\left(\frac{.44}{.52}\right) = -.24.$$

Now the Yes-designation for $(\alpha)_1$ is revealed, and both B and C eliminate the wine-areas where $(\alpha)_1$ does not apply. A then presents the second aspect, eliciting assessments of .20 and .40, and then the third, eliciting .78 and .30. The action is summarized by a protocol of play:

Aspect	Truth-value	Assessments B	Assessments C	Cumulated Score
$(\alpha)_1$	Yes	.44	.52	-.24
$(\alpha)_2$	No	.20	.40	.16
$(\alpha)_3$	Yes	.78	.30	1.52

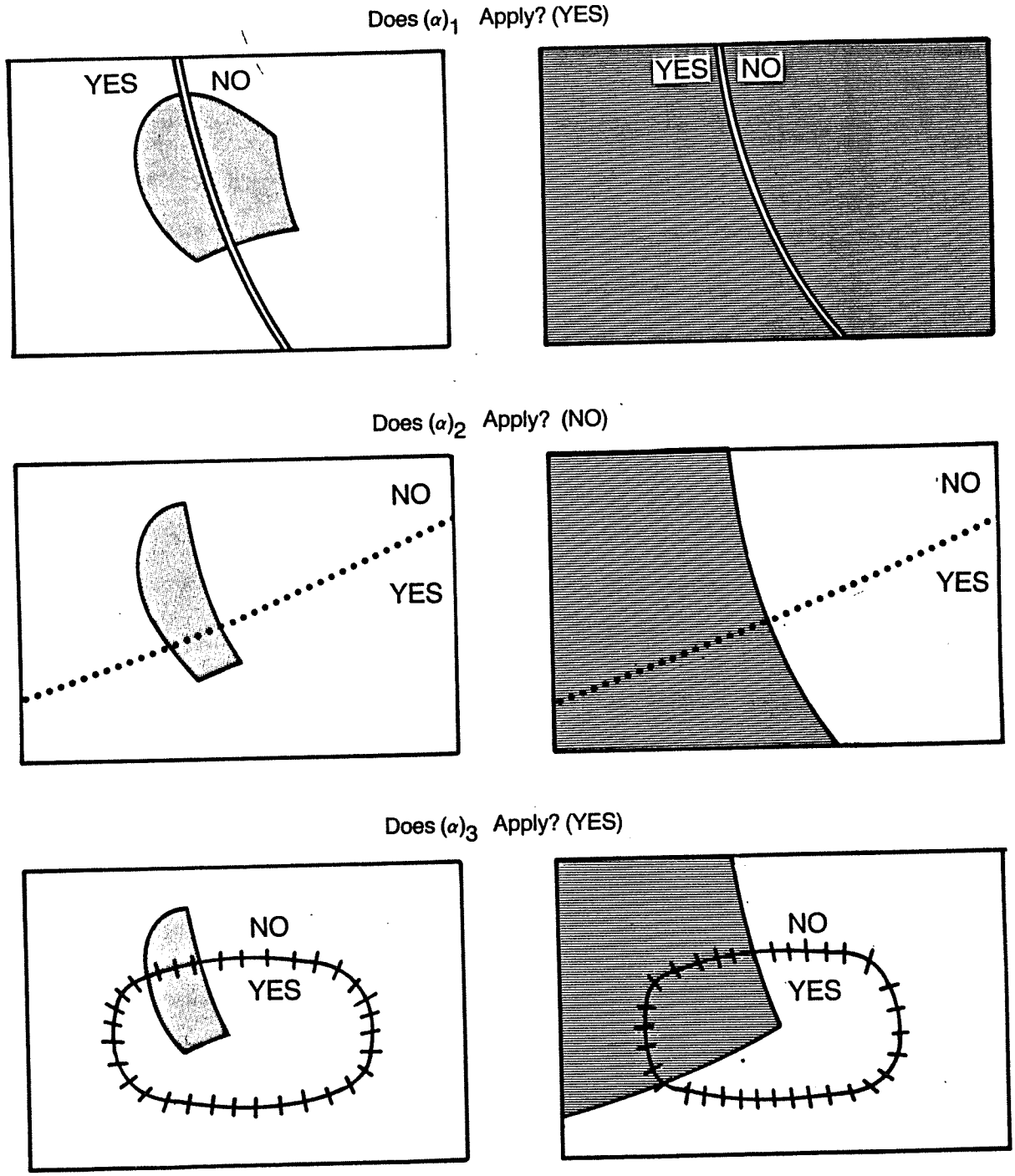


Figure 13: Shaded areas indicate subsets of "wine-universe" that B (left panels) and C (right panels) consider possible, before aspects 1 (top), 2 (middle), and 3 (bottom) are revealed by A.

The final score, +1.52 in favor of A and B, is what we would have obtained in the basic (one-stage) game, in which A presents all eight possibilities, and B and C assign probabilities $p(\alpha_{101}|\beta_{101}) = .270$ and $p(\alpha_{101}) = .094$, respectively.

Proposition 5 assures us that all ways of "zeroing in" on the correct observation are equally good. The Table below simulates play when A has adopted the strategy of going through the eight possible observations in sequence, and asking whether each is the correct one (we denote the single-observation aspects $\{\alpha_{ijk}\}$ by $(\alpha)_{ijk}$):

	Aspect	Truth-value	Assessments B	Assessments C	Cumulated Score
1.	$(\alpha)_{000}$	No	.11	.16	.07
2.	$(\alpha)_{001}$	No	.31	.06	-.38
3.	$(\alpha)_{010}$	No	.00	.25	.04
4.	$(\alpha)_{011}$	No	.29	.13	-.25
5.	$(\alpha)_{100}$	No	.17	.42	.27
6.	$(\alpha)_{101}$	Yes	.75	.31	1.52
7.	$(\alpha)_{110}$	No	.00	.00	1.52
8.	$(\alpha)_{111}$	No	.00	.00	1.52

The final score is +1.52, again. (The game could have ended, in fact, after the Yes-designation on $(\alpha)_{101}$). The significance of Proposition 5, then, is that there is really no strategy to playing the sequential game -- one way of fully identifying an observation is just as good as any other.

8 DESIGNING AN INFORMATION-PUMP: AN EXPERIMENTAL EXERCISE

We have conducted the discussion and analysis of these games under the assumptions that the participating players are perfect Bayesian agents, with a fully developed and mutually consistent structure of subjective probabilities, that they have no motive in play except to maximize their individual expected score, and that all this is common knowledge among them. There will be a gap, clearly, between this ideal conception and the play of actual subjects, no matter how well instructed and motivated. What is less clear, however, is whether the strategies employed by experimental subjects will have sufficient resemblence to theoretical prediction to make

the game a useful instrument for collecting introspective judgements, or -- as is quite possible -- that the strategic equilibrium which we have derived theoretically is a fragile one, and will fall apart under the imperfect conditions of actual play.

Given such concerns, it seemed appropriate to complete this theoretical article with a brief description of an experimental procedure that implements the elicitation-by-aspects version of the game, and a survey of the results that have been obtained with it so far. Altogether, we used seven distinct series of stimuli, which were selected with two basic questions in mind:

(1) Do final scores accurately reflect the mutual information shared by A and B?

(2) Are the aspects defined by A "introspectively" reasonable?

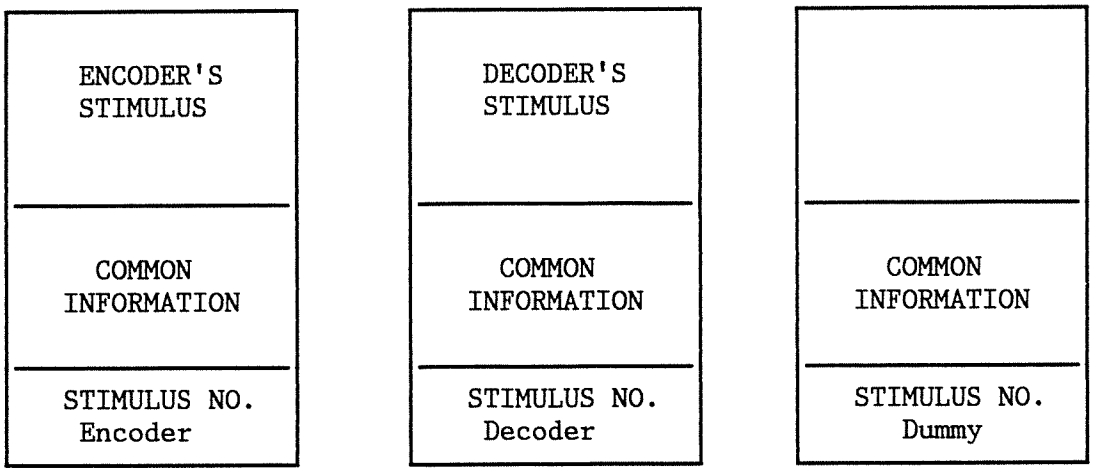
The angle of approach to these questions was somewhat different for different groups of stimuli, which is what I would like to briefly explain. The first two series made use of color chips, and names of birds, thus giving us samples of play with two very different sources of mutual knowledge, phenomenal and semantic. In series 3 through 6, I attempted to vary the information available to Player C, holding the mutual information of A and B constant. Specifically, in all four series, the stimulus shown to A and B was a single Chinese character; in Series 3, 4 and 5, however, C also saw this character, but concealed as an element of 2, 4 or 8 possible stimulus characters. The amount of mutual information available to A and B was at most 1, 2, or 3 binary units, corresponding to the 2, 4, or 8 possible characters presented to C. This manipulation created a numerical upper bound on the score that A and B should have been able to attain.

Finally, with Series 7 we wanted to elicit specifically relational features. In this study, each subject was exposed to distinctive,

personalized versions of artificially constructed stimuli, and was never given the opportunity to examine the versions available to the other subjects. The different versions were related, however, in being different pictorial representations of the same random matrix (specifically, a random drawing of five cards from a Poker deck). In this way, even though the concrete stimuli shown to the any of subjects playing roles A and B invariably looked very different (see Figure 18), the subjects did nevertheless possess a latent mutual knowledge of those aspects of each other's observations that were associated with higher-order relations of stimulus resemblance, category, and the like, and that were more-or-less invariant accross the different stimulus versions. To be sure, they could not have been aware of this mutual knowledge at the start of the experiment; a secondary question for this experiment, then, was to see whether subjects could recognize the mutually meaningful relations by playing the game.

Procedure Subjects played the game from three separate rooms, by communicating through inter-connected computer terminals. A session always consisted of six rounds of play, in which the roles, A,B,C, were permuted among the subjects in the six possible ways. In any round, Player A was called the Encoder, Player B the Decoder, and Player C the Dummy.

At the start of the session, each subject received a booklet containing six personal information sheets, one for each round of play. The information sheet was divided horizontally into three parts like this:



Player A's information sheet | Player B's information sheet | Player C's information sheet

The bottom part identified the stimulus number, and informed the subject of his role for the round. The middle part provided some information about the stimulus used for the round; this common part of the information sheet was physically identical in all three booklets. The stimulus itself was located in the top part of the information sheet; for Player C, this area was always blank.

Player A initiated play, by entering a string of letters, called a feature, and designating it, with a YES or NO response, as appropriate or inappropriate characterization of the stimulus. The program next requested a second feature, and so on, until the specified number of features (either 6 or 12) was supplied. Player A had 60 seconds to type in the first feature, and 30 seconds for each subsequent one.

The scoring As soon as a feature was entered, it was communicated to Players B and C, who assessed a probability (expressed as percentage, on a 0-100 scale) that the feature bears a YES-designation. The true YES/NO value was then revealed, and the person's assessment score calculated by

the logarithmic scoring rule, normalized so that a response of 50, indicating complete ignorance, received a score of 0, and a correct maximum confidence response of 100 or 0 received a score of +100. (The program actually interpreted these responses as 99.9 and 0.1, respectively, to avoid the infinite penalty associated with a 0/100 mistake). So, when B or C entered a probability, x, for a YES-designation, their assessment score could be one of two values,

$$100(1+\log_2(x)), \quad \text{or,} \quad 100(1+\log_2(1-x)),$$

depending on whether A had selected YES or NO as the correct designation of that feature. If no probability was entered in 30 seconds, the program would interpret that as a response of 50. When assessing a probability for the second, and subsequent feature, Players B and C would see on the screen a record of the previous features, the correct YES/NO values, and their own assessment scores; however, the net score, which is the cumulated difference of B's and C's assessment scores, would only be revealed at the end of the round. After the session, the subjects were informed of their cumulated individual scores for the six rounds, as well as which among them did best in the roles of A, B, and C. For stimulus series 3-6 the scores were used to calculate individual bonuses, which added (on average) 30% to the $5/hour rate of pay; for the other series, they were informational feedback, only.

Subjects The subjects were undergraduate students from Harvard College. They received an instruction booklet which described the rules of the game, reproduced a short protocol of hypothetical play, and gave some hints about strategy. All subjects had some training in the procedure

(with unrelated stimuli) prior to experimental play, but this varied from one to several sessions. One subject did not approach the task responsibly, and we excluded data from sessions in which he participated.

Summary of Scores The best that A can do in this game, is to select features which, at each stage, divide the remaining observational possibilities into two roughly equiprobable sets, so that C is forced to assess a one-half probability, and score zero. If B can track these same features with 100% confidence, then the net score will accumulate at 100 points per feature. So, even if Player C is completely in the dark about the correct feature designation, he or she can still limit the loss to 100 points per feature, by consistantly assessing a probability of .5. The number of features, per round, times 100, is an upper bound that scores should rarely exceed.

Table 2 identifies the different stimulus series, and summarizes the results. As we can see from the median scores, and the relative frequency of wins, the informed pair of subjects had a robust advantage in the game.[17] In ideal circumstances, we would be able to interpret the average score in each condition (divided by 100) as an estimate of the quantity of mutual information, in binary units or bits, between the observations of A and B. So, for example, an average score of 358 indicates 3.58 bits of

[17] One might note that C's chance of winning are not negligible, even when he knows very little. Consider, for example, a situation where B knows the correct feature designation 99% of the time, while C is completely ignorant. Over a 6-feature round, one would expect B able to push this advantage, and virtually always ensure a win. This is not so. In order to win, B must not make a single mistake, since the penalty for a .99 error is more than five times the gain for a .99 correct response; the probability of 6 consecutive correct guesses is $(.99)^6$ or .94. This still leaves C a 6% chance of winning the round.

information, or equivalently, perfect mutual discrimination, by A and B, of about 12 (= $2^{3.58}$) equiprobable observations.

Recalibration A problem, however, with such a direct interpretation of scores, is that the subjects' assessments confirmed a previously well-documented pattern of over-confidence (Alpert and Raiffa, 1968; Lichtenstein et. al., 1982). Subjects would typically be wrong about 5-10% of the time even when they claimed absolute certainty (i.e. 0 or 100 response), and this bias persisted at the intermediate probability levels.[18] Overconfidence is a form of assessment error, and will reduce the average individual scores of both Players B and C. Unfortunately, the net effect of this error need not be zero, since Player B normally assesses probabilities closer to the extremes, and it is there that the penalty for mistakes is quite spectacular. So, a uniform tendency to overconfidence on part of B and C may produce scores that underestimate the mutual information created by observing the stimulus.

To the extent that we wish to attend to, and draw inferences from the numerical scores, we have a choice of either relying on non-parametric statistics, like the median, which suppress the impact of negative outliers, or we can attempt to "recalibrate" the subjects' probability assessments by shading them systematically towards one-half. In conjunction with this series of experiments, we developed an automatic recalibration program for calculating personal recalibration coefficients, δ, which are applied as exponents to subjects' assessed odds of YES/NO designation:

---

[18] A subject is over-confident if his or her assessments of probability x (greater than .5) have higher than (1-x) probability of being wrong. Similarly, a subject could be under-confident if the assessments are wrong less than (1-x) fraction of times.

Series	Stimulus Type	Common Information	No. of Teams	Features/ Stimulus	No. of Rounds	Median Score	Proportion of Rounds won by A&B
1	Color Chip	Trade name of color	1	6	18	+1.79	.89
2	Name of a bird	none	1	12	12	+4.80	.75
3	Chinese Character	2 possible characters	3	6	6	+1.32	.72
4	"	4 possible characters	3	6	6	+1.92	.78
5	"	8 possible characters	3	6	6	+3.31	.72
6	"	none	3	6	6	+5.16	.83
7	Pictorial Representation of 5-card hand	none	1	6	60	+ .96	.70

Table 2: Summary of experimental conditions.

$$(\text{recalibrated odds}) = (\text{assessed odds})^{\delta}.$$

Thus, a subject with a coefficient of .5, would have a .9 assessment (i.e. 9:1 odds) interpreted as a .75 assessment (i.e., 3:1 odds). The value of each subject's coefficient was the one that optimally adjusted his assessed probabilities, in that his or her total personal score across all sessions was maximally improved by the transformation.[19]

The median recalibration coefficient, for the 16 subjects, was .62, with the lowest value being .29, and the highest 1.37 (indicating underconfidence). The Table, below, shows the effect of recalibration on selected probabilities, for these three coefficients:

	Recalibrated Probability		
Assessed probability	δ=.29 (lowest)	δ=.62 (median)	δ=1.37 (highest)
Certainty (.999)	.88	.99	.999
.98	.76	.92	.99
.90	.65	.80	.95
.75	.58	.66	.82
.40	.47	.44	.36
.05	.30	.14	.02

---

[19] Specifically, if x_i is a subject's assessed probability that aspect "i" bears a YES-designation, and F_Y and F_N are, respectively, the sets of all YES- and NO-designated aspects that this subject encountered, then the subject's optimal recalibration coefficient is the value of δ that maximizes the gain in personal score, $G(\delta)$:

$$G(\delta) = \sum_{i \in F_Y} \log \frac{x_i^{\delta-1}}{(1-x_i)^{\delta} + x_i^{\delta}} + \sum_{i \in F_N} \log \frac{(1-x_i)^{\delta-1}}{(1-x_i)^{\delta} + x_i^{\delta}} .$$

In our interpretation of the results we will make use of both raw and recalibrated probabilities; except for Study 2, the conclusions we wish to draw from the data are primarily qualitative, and are not affected by recalibration. I bring this problem to attention, however, because it will crop up in any future attempt to implement the procedure, and one will need to have ready some recalibration technique to neutralize it.

Study 1: Birds and Colors

The first two series of stimuli were introduced to see how subjects would play with a perceptual, and a semantic structure of mutual knowledge. The stimuli in the series 1 were 18 color samples (chips) for commercial wall paint, produced by Company X. The trade name for each color was written on the common part of the information sheet, thus providing the uninformed player with a potentially important clue. In the second series, A and B saw the name of a bird chosen from the set of twelve that Malt and Smith used to collect feature lists (cited in Smith and Medin, 1980, pp. 103-105). C had no information, except that the "stimulus" had to be a bird. Player A was instructed to restrict himself to aspects that would be relevant even if the English word for the bird happened to be different from the actual one; in this way we wanted to exclude references to letters contained in the word, rhymes, and other "non-semantic" aspects of the stimulus.

Let us start by looking at some sample rounds. Figure 15 introduces a convenient method for representing how play evolves over the course of a

Insert Figure 15 About Here

round. The two lines in each panel are plots of cumulated uncertainties, by feature number, of B and C. (The uncertainty cumulated up to feature n is the sum,

$$\sum_{i=1}^{n} \log(1/x_i),$$

where x_i is the probability assessed for the correct Yes/No designation.) The plot for a person who is certain of the correct designation every time would never leave the horizontal axis; in contrast, the plot of a person whose assessments are at chance would follow along the diagonal line in the Figure, which increases at a rate of 1 bit per feature. The gap between C's and B's cumulated uncertainties is then the net score, cumulated by the AB team, up to that feature number.

The two panels in Figure 14 give two very different examples of play. In the top panel, A and B managed to add a net point per feature, whereas

Insert Figure 14 About Here

Insert Table 2 About Here

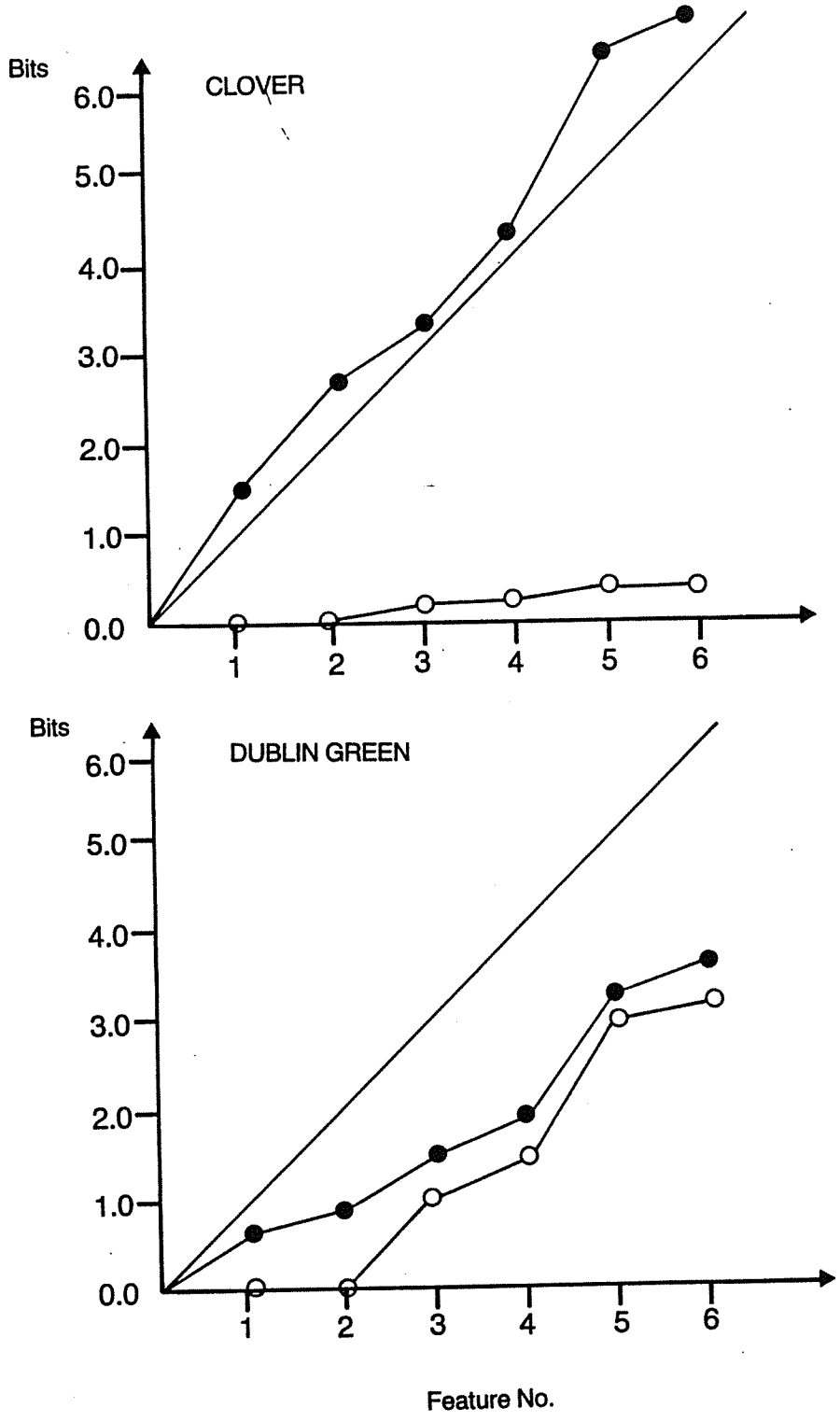


Figure 14: Examples of play with an easy (Clover) and hard (Dublin Green) color. Open and solid circles are cumulated uncertainties for B and C.

the final score in the bottom panel was essentially zero. The probable reason for this was the difference in the amount of information that the color name conveyed to C before play began. The top-panel color was "Clover," whose name derives from the (pinkish) color of the clover flower, and not from the leaf, as one might think. The first aspect encoded by A took advantage of this equivocation, as we can see in the Table below:

"Clover" Protocol

Aspect (* = yes)	Assessment B	Assessment C	Cumulated Net Score	Recalibrated Net Score
lighter shade of grass	.01	.65	1.50	1.56
*light	.99	.45	2.64	2.70
*floral prints	.90	.65	3.11	3.01
blueish	.01	.50	4.10	3.97
common lip color	.10	.75	5.95	5.82
*the traditional color for a baby girl	.99	.75	6.36	6.16

By the fifth aspect, C knew that the color was neither green nor blue, was light, and could appear on floral prints; the high response to "common lip color" suggests that he places the color in the red-pink region, but for that aspect the information actually misleads.

For the color in the bottom panel, the name was a complete give-away, as most residents of the Boston area have a pretty good idea of what "Dublin Green" is like. The challenge for A, in this case, was to find aspects that will extract the small amount of additional information provided by the actual color chip. The rise of B's cumulated assessment line above the horizontal indicates that A has introduced aspects that are quite difficult to assess. Here is the protocol of play:

"Dublin Green" Protocol

Aspect (* = yes)	Assessment B	Assessment C	Cumulated Net Score	Recalibrated Net Score
very yellowish	.01	.35	.61	.49
*a Celtics fan might wear this color	.99	.85	.83	.60
like pine needles	.50	.35	.45	.18
a soothing color for walls	.25	.25	.45	-.04
*grass	.35	.40	.26	.03
*vibrant	.89	.79	.51	.05

The bird stimuli presented a somewhat different encoding problem for Player A. Whereas no one aspect was likely to exhaust the informational content of a color chip, the identity of the bird presented to A and B could certainly be disclosed with a single clue. Therefore, Player A had to select the initial aspects with care, lest they prematurely identify the bird name, and so equalize the expected assessment accuracy of B and C over the remaining aspects.

This is indeed what happened for several of the stimuli. Figure 16 gives two examples where in which we can spot the aspect at which C (probably) identified the bird. Here is a reconstruction of the "Falcon" protocol (plotted in the top half of Figure 15):

Insert Figure 15 About Here

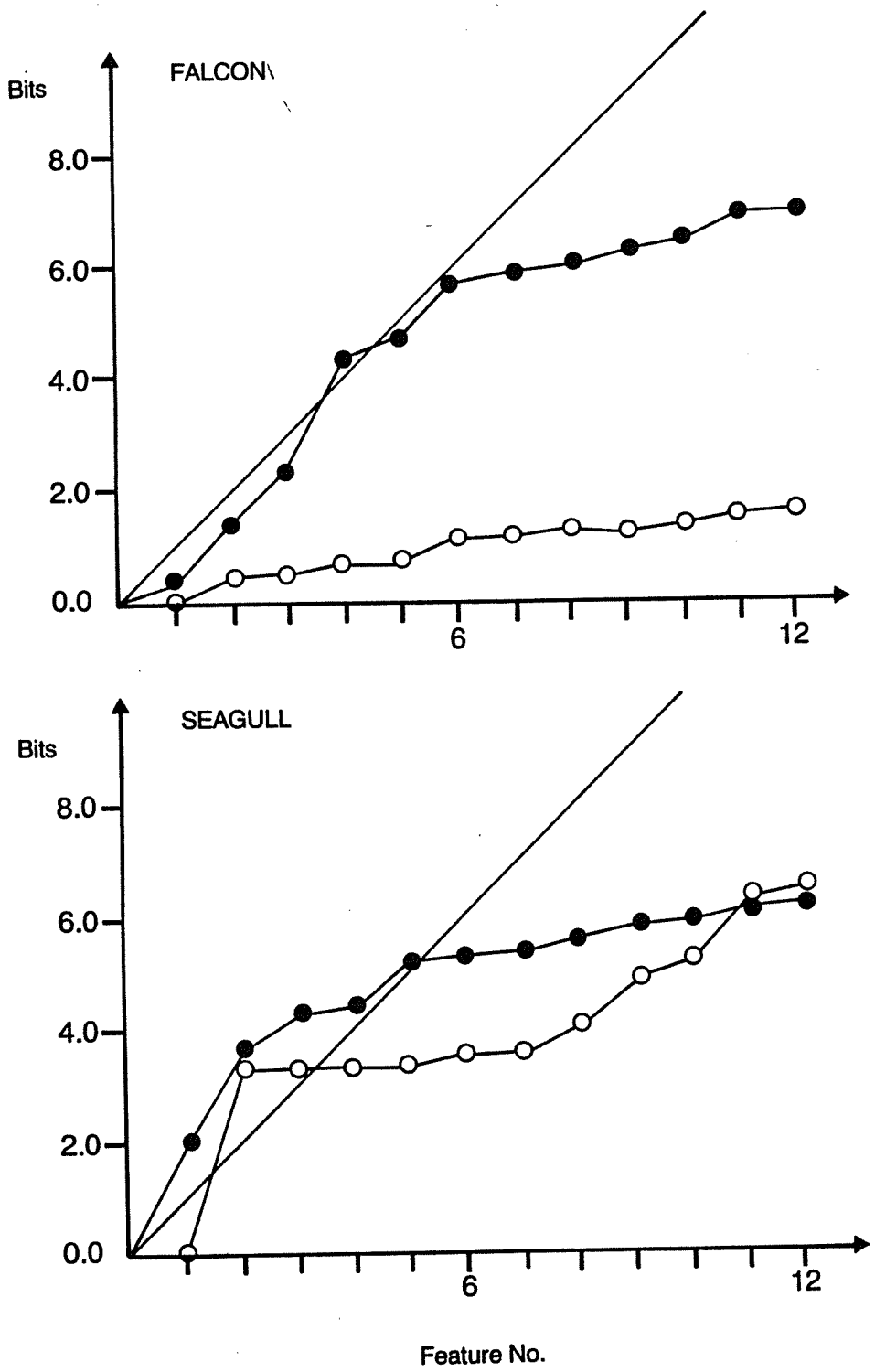


Figure 15: Two examples of premature identification of bird-name (see text). Open and solid circles are cumulated uncertainties for B and C.

"Falcon" Protocol

Aspect (* = Yes)	Assessment B	Assessment C	Cumulated Net Score	Recalibrated Net Score
*can fly	.99	.75	.41	.34
*lives in Europe or Africa	.75	.50	.99	.80
1960's symbol of popular discontent	.05	.50	1.34	1.01
*eats meat	.90	.25	3.19	2.86
*has sharp beak	.95	.85	3.35	2.89
*domesticable	.80	.50	4.03	3.44
*used more by nobility	.95	.85	4.19	3.47
*sharp claws	.95	.90	4.27	3.42
kills elephants	.01	.15	4.49	3.58
kills cows and deer	.05	.15	4.65	3.61
eats lettuce	.15	.25	4.84	3.63
*kills small animals	.95	.95	4.84	3.52

Player A begins by excluding non-flying birds, (which according to C have a 1/4 chance of appearing). The next aspect places the bird outside of the United States -- a significant restriction. Player C is somewhat surprised by "eats meat," but not, subsequently, by the "sharp beak." It is the aspect, "domesticable," that triggers the critical insight, as we can see from the sharp upturn in C's cumulated score, in Figure 16. After that, B can squeeze out only a slight advantage over the next six features, and that advantage disappears after recalibration.

The protocol for "Seagull" tells a similar story (bottom panel in Figure 15).

"Seagull" Protocol

Aspect (* = Yes)	Assessments B	Assessments C	Cumulated Net Score	Recalibrated Net Score
*white	.99	.23	2.11	1.52
*more than a foot in height	.10	.36	.25	.21
found in Australia only	.01	.36	.76	.86
does not eat fish	.01	.10	.90	1.20
cannot fly	.01	.40	1.63	2.00
often shot by sportsmen	.15	.13	1.59	2.07
*considered a pest by many	.99	.96	1.63	2.25
can be domesticated and raised for food	.25	.12	1.40	2.13
*popular iconographic figure	.55	.83	.80	1.73
bred in zoos	.15	.04	.63	1.59
cannot swim	.45	.15	.00	1.17
has a pleasant song	.15	.02	-.18	.96

After three aspects, C is thinking about a large, white bird, not restricted to Australia. Although he is confident that the bird eats fish, he is still not entirely sure of its identity, as the tentative judgment on the next aspect, "cannot fly," proves. (At this point, C may have narrowed down the stimulus to either a seagull or a swan, both of which are relatively large, white, fish-eaters.) Adding "flies" to the existing inventory of aspects dispels all mystery, and from that point on all of C's assessments are confident, and correct.

The mean net score for colors and birds, across all rounds, were 2.08 and 4.73 points, respectively. For the bird series, in particular, it is tempting to relate this to the number of possible bird names. Can we claim to have obtained an estimate of this number, as in the neighborhood of $2^{4.73}$, or 26 (equiprobable) birds?

Theoretically, we can justify such inference with the following argument. Unlike the colors, the bird stimuli give rise to a structure of mutual knowledge, that might be called categorical, in that the mutual information between observations α and β, given that the stimulus belongs

to a particular category (i.e., a specific bird name) is zero. Once C intuits the category-membership, the advantage of Player B disappears completely. With such structures, the mutual information, and hence the expected score in the game, is bounded by the uncertainty in category membership.[20] Thus, if the number of possible bird names is somewhere between 16 and 64, then the score in the game should not exceed 4-6 points, even with perfect play. In practice, our confidence in the actual score as a measure of this upper limit depends critically on the ability of Player C to incorporate revealed information into his new assessment. It is not clear that C will be able to do this, especially since he has no more time than B to assess a probability. Although in theory both B and C have to update their beliefs on the basis of A's feature-designations, in practice the updating the B has to do in quite minimal, confined perhaps to those occasions when there is some significant confusion between him and A; C's updating, in contrast is essential to his success, and to the success of our procedure, generally. What is at issue, here, is not just the interpretation of average score, but the ability of the game to extract a full description of A's observations: If A can score by recycling or recombining old features, then he will not be under pressure to produce an exhaustive inventory.

---

[20]Formally, if the category membership is given by variable σ, and if $I(\alpha:\beta|\sigma) = 0$ (α and β are independent, given σ), then

$$I(\alpha:\beta) \leq I(\alpha:\sigma) \leq U(\sigma).$$

Study 2: Do scores measure information?

The second study was conducted to test whether the scores would observe the theoretical limit, when that limit is artificially created. The stimulus, throughout, was a standard Chinese character; the common information--that is, the information made available to C as well, was a set of 2, 4, or 8 characters which included the stimulus. For comparison purposes we added a fourth condition, in which there was no common information. In the first three conditions, then, we induced a categorical structure of mutual knowledge, with 2, 4, or 8 possible categories. The average scores in these three conditions should not exceed 100, 200, or 300 points respectively.

Figure 16 plots median and mean scores (both raw and recalibrated) as a function of the theoretical maximum. (For the fourth condition, in which C had no clues, the theoretical maximum is just the 600 points that can be accumulated over six rounds.) Both means and medians are indeed below, or just a the theoretical maximum. The differences in means of pairs of conditions are significant (by the t-test, at the .05 level) for 2 versus 8 alternatives ($t = 1.86$), and for 8 alternatives versus no common information ($t = 2.15$), but not for the 2 - 4, nor the 4 - 8 comparison.

Insert Figure 16 About Here

Figure 17 plots separately the cumulated uncertainties of Players B and C, from which we can get a fuller picture of how stimulus set size affects play. From the bottom two panels, we see that the difference in score between the 8-alternative and the no-restriction conditions derives entirely from a change in C's assessment score. When the stimulus is chosen from a set of 8 possibilities (bottom-left panel), C's uncertainty starts to level as soon as the YES/NO value for the first two features are

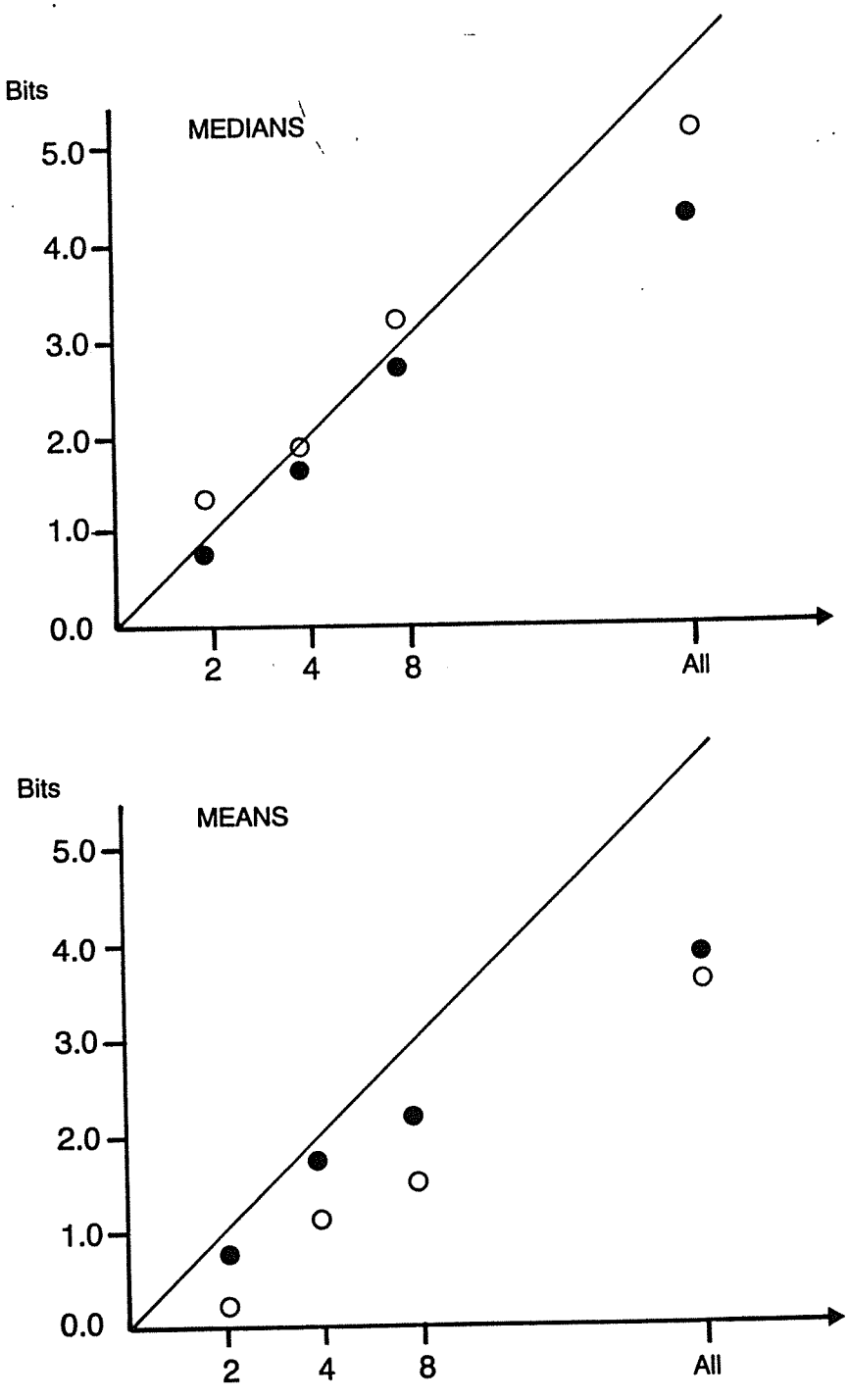


Figure 16: Mean and median net scores in Series 3-6, plotted as a function of common information. Open and solid circles designate raw and recalibrated scores, respectively.

revealed; in contrast, when there are no clues (bottom-right panel), C's guesses remain at chance for the entire round.

Insert Figure 17 About Here

In the top two panels, we see a more gradual separation of B's and C's uncertainties. It is interesting to speculate why this is so. Generally, subjects found the two-alternative condition the most difficult one to play, especially in the role of A. In order to avoid revealing the stimulus with a single feature, they would initially select features whose YES/NO designation conveyed little information to C. One class of such features were those that could be construed to apply (or not apply) to both stimulus possibilities, in which case C's assessment accuracy would rise above the chance level; another class were very unusual, ambiguous descriptions, which even B had trouble assessing with confidence. The use of these two types of features contributed to a lower assessment score for B, and a higher one for C (relative to the 8-alternative, and the no-clues conditions). However, the strategy of gradually leaking information did not yield any real benefit, as we can see from the top two panels in Figure 21. Player A could just as well have identified the stimulus with the first feature, in the two-alternative condition, or with the first two features, in the four-alternative condition, and collected the 100, or 200 points in net score.

Study 3: Stimulus Relations

The third study was conducted in a somewhat more speculative frame of mind. Our intention was to present subjects with stimuli that would initially resist description, and so encourage them to uncover invariant relationships among pairs, or groups of stimuli. To do this, we developed

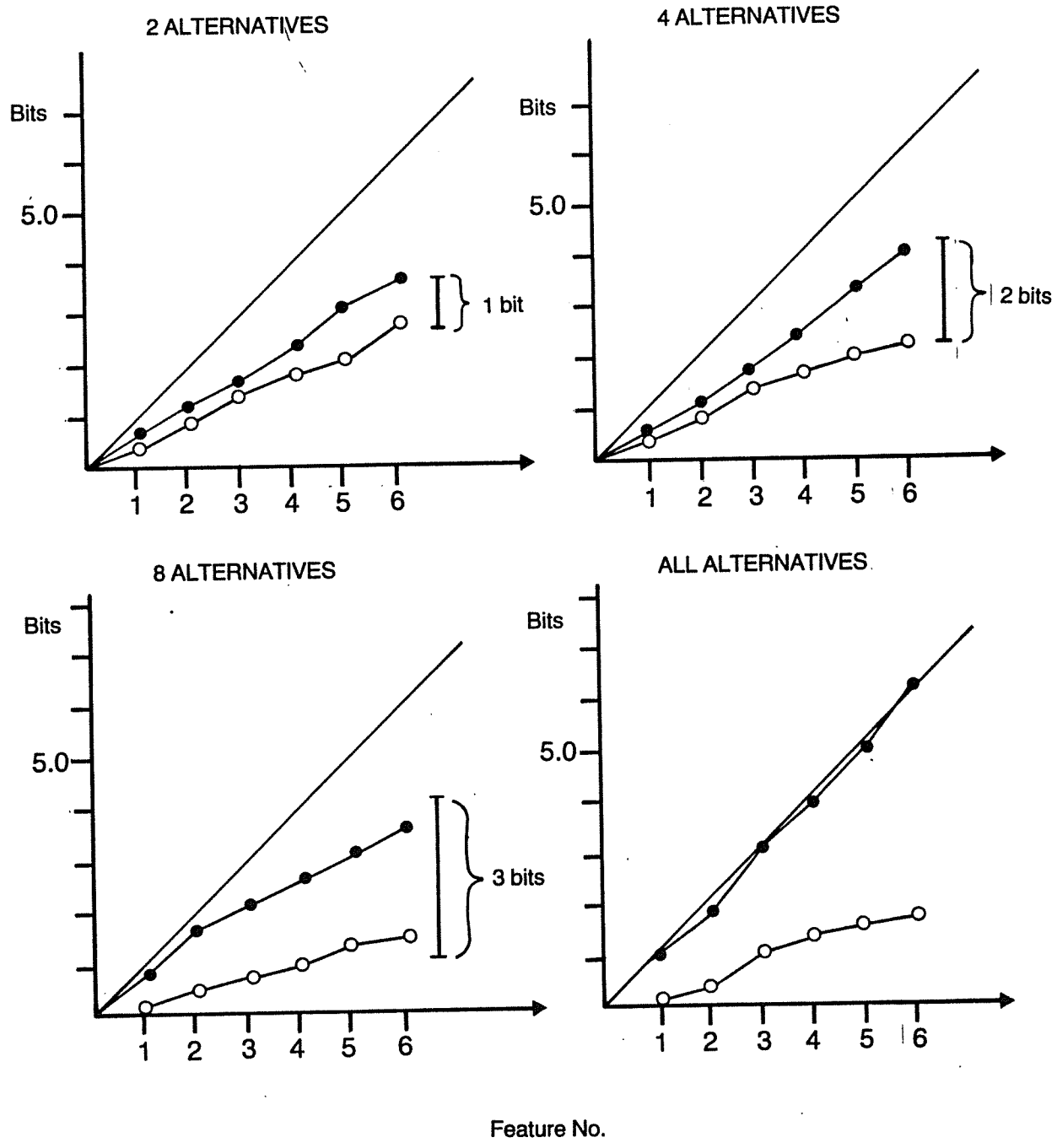


Figure 17: Mean cumulated uncertainties (recalibrated), by feature, in Series 3-6. Open and solid circles indicate Players B and C, respectively.

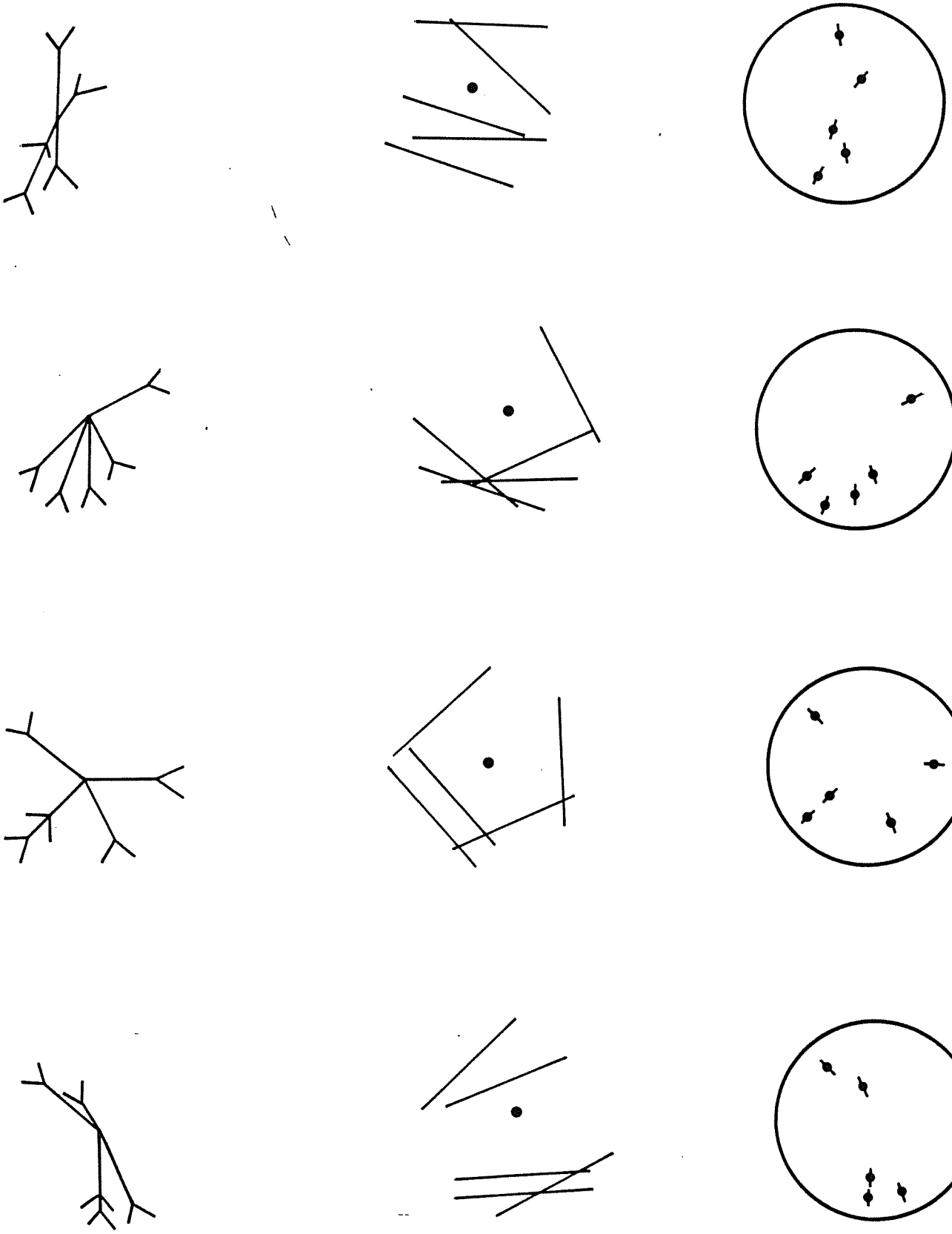
three distinct sets of stimuli, each of which could be interpreted as an abstract representation of a five-card Poker hand. Throughout the experiment, each subject saw only samples from one of the three representations, but on any given trial, the stimuli shown to A and B would be representations of the same Poker hand.

The stimuli were constructed in the following way. Each card in the 52-card deck had an assigned location within a circle of radius x. In polar coordinates, the suit determined the radius -- i.e. distance of the location from the center of the circle, while the card value (A, K, J,..., 2) determined the angle relative to the top of the circle. A five card hand thus created a matrix of five points in the circle, which was then embellished, as it were, in three different ways, producing a matched triple of stimuli.

We can refer to these three types as roots, sticks, and boats. A root was made by drawing a line from the origin to each of the five points, and decorating the end with a small v-shaped fork. A stick was made by adding a dot at the circle center, and drawing, at each point, a straight line perpendicular to the center. Finally, a boat, was made by enlarging each point to a dot, and cutting it with a short line, oriented towards the center of the circle (which was here explicitly shown).

Figure 18 shows four matching triples of stimuli. One can see that there are some interesting correspondences (and lack of correspondences) in both local detail and global organization. For example, the top and bottom triple are quite similar - especially in the sticks version. The two top triples look different overall, but share a common detail, which was

Insert Figure 18 About Here



Four examples of matching root-stick-boat triples.

created by a pair of same-value cards in the hand. This aspect shows up as a line with two forks in the roots version, a pair of parallel lines in sticks version, and a pair of dots that point at each other in the boats version. In contrast, there are other aspects that are salient in only one version; for example, the appearance of sticks is strongly affected by line intersection which has no clear counterpart in boats and roots.

The subjects were of course not aware of these perceptual relationships. Before the start of the experiment, they were each given a similarity-judgment exercise, to familiarize them with their own types of stimuli. A short cover statement explained that in this experiment, each subject would see a "personal version" of the stimuli, and that "it may be helpful to think of the different versions as something like different traces or representations of a single object or process." Subjects were also enjoined not to discuss the stimuli, for any such discussion, they were told, would invalidate the results of the experiment.

The first two rounds of play yielded a completely negative result: Subjects could make no sense of each other's descriptions, and to that extent found the procedure quite frustrating, One person's references to properties of branches, or angles, or point groupings, meant nothing to the other two, as we can see from the first two points in Figure 19.

Insert Figure 19 About Here

Starting with the third round, each subject received a reference set of twelve drawings, which were identified by letter, and which were matched across the three sets so that drawing "F," for example, represented the same five-card hand in all three versions. This allowed the subjects to describe the stimulus in purely relative terms, as similar to, or belonging with, or standing in some other abstract relation to specific drawings in

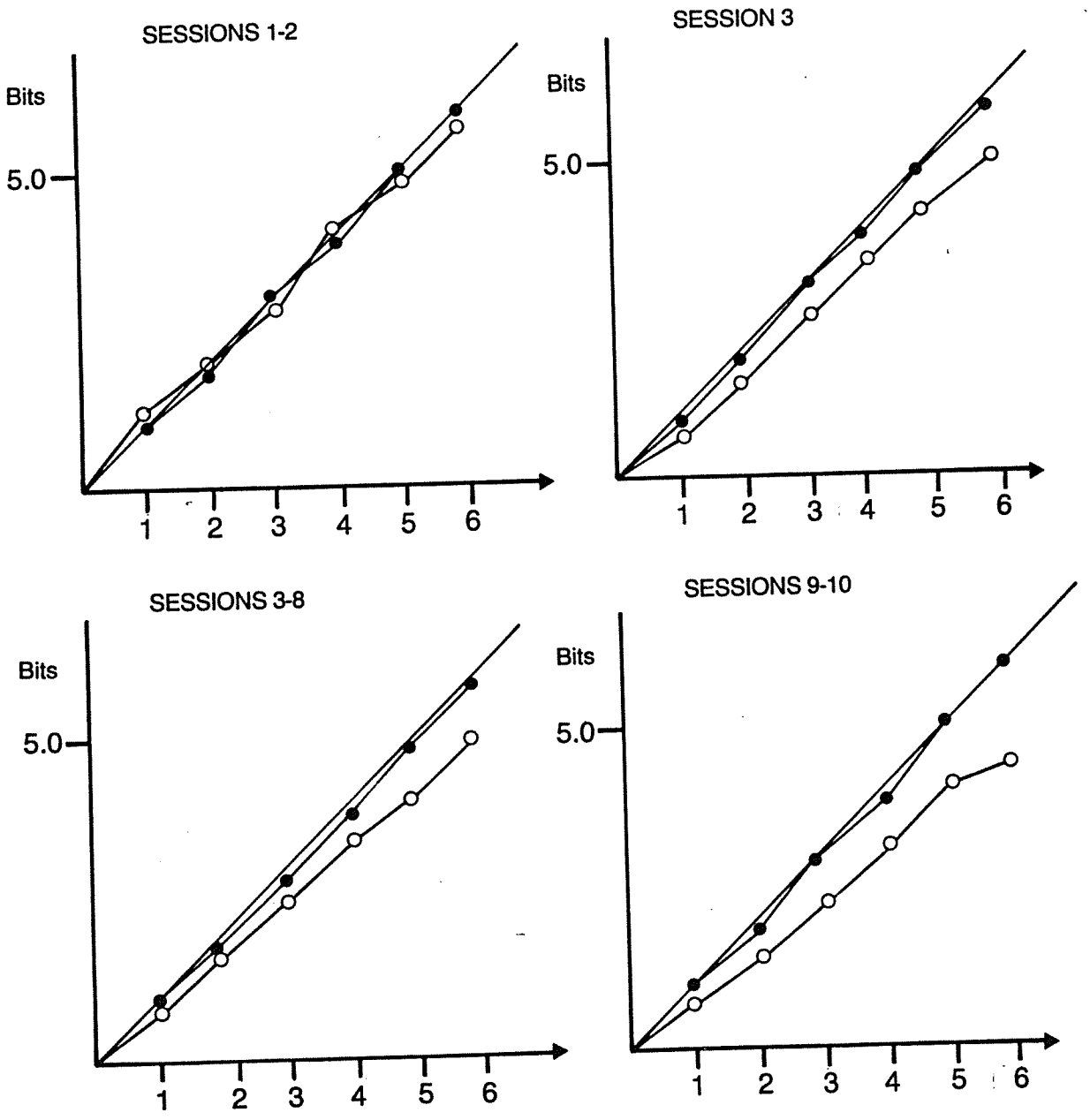


Figure 19: Mean cumulated uncertainties for B (open) and C (solid), in Series 7. Session 3 was the first one in which players A and B had a reference set available. The set was withdrawn for sessions 9 and 10.

the set of twelve. Subjects were not required to make use of the reference set, but it was expected that they would do so, given their lack of success in the previous two rounds.

This is indeed what happened. The proportion of relational features, that is, features which explicitly referred by letter to one or more reference drawings, climbed from an initial 47% in Round 3, to a high of 77% in Round 6, and then declined to 55% in Round 8, the last round in which the reference set was available. As we can see from Figure 19, there was an immediate separation in B's and C's assessment scores, indicating some agreement over stimulus relations.

In the last two rounds of the experiment, the reference sets were withdrawn, and the subjects were thus returned to the same procedure they played in rounds one and two. Although Player A could no longer use relational features, this did not lead to any loss in net score -- if anything, the score improved, with respect to the previous six rounds.[21]

---

[21]After the study, I asked the subjects to describe briefly what they thought the other stimuli looked like. The most accurate answer was given by the "boats" person, who guessed that one version had "lines connecting the dots to the center point," and the other "had lines drawn to the sides of the circle, going at right angles to the axis lines through my dots." This was incorrect only in that it ommitted the "forks" at the ends of the root-lines, and supplied a non-existent circle as the boundary at which the lines in the sticks-version stopped. The person with the roots version had a clear conception of boats -- as "the points themselves, without connections to the center," but was confused about sticks. She made an interesting, but wrong hypothesis that the lines which this person was referring to were created by connecting the points to each other in some way. The person with the sticks-version only knew that intersection of lines was not a "reliable clue," and speculated that other people might have the lines expanded or rotated in some way to account for this.

Are the features introspectively reasonable?

In reviewing the thousand-plus features collected so far, one easily forgets that the subjects had no explicit instructions about the sort of descriptions they were expected to generate. The scoring rule imposed two purely informational criteria, which were, first, that the features had to focus on the mutual information created by the stimulus, and not by the commonalities of the experimental situation, and, second, that the features should be non-redundant, that is, they should not repeat information supplied by earlier features. Beyond these two conditions, which were enforced not by the experimenter but by the person who happened to be playing the role of C, subjects had complete freedom to choose whatever statements they thought would work.

Thus uninstructed, subjects produced very heterogeneous lists of features, combining ordinary and obvious items with others that were quite original. Among the things found in Chinese characters, for example, were: "a cat," "fishbones," "glasses," "dancers," "a kangaroo," "R2-D2," "gumby," "an eye," "a face," "Lincoln's hat," "two people talking to each other," "an intense game of tic-tac-toe," "a coyote baying at the moon," etc.. Such frankly projective features were interspersed among geometrical/topological descriptions -- counts of lines and dashes, separation into parts, symmetry, horizontal/vertical organization, complexity, and the like.

Subjects often "engaged" the stimulus in an active way, noticing what could be done with it, how it would fit in a certain context, etc. Here is a sample of such features:

Stimulus Type	Feature
Chinese Character	...could hold water if turned upside down.
" "	...would rock easily on its base, it seems.
Bird (chicken)	...is not used on heraldry crests.
" (starling)	...most people have not seen one.
Color (Deep Lagoon)	...a trendy color for clothes.
" (Minaret)	...would make an attractive tan.

The relational aspects collected from Series 7 further reinforce this impression of an active attitude. Subjects made us of many different kinds of stimulus relations, such as (letters identify stimuli in the reference set):

Relation	Example
Similarity	Much like J.
Relative similarity	More like L than like H.
Conditional similarity	Top/bottom distribution like F.
Categorization	Most like I,M,E,J.
Abstraction	Has feature in common with E and H.
Transformation	Is a rotation of A.

The general impression one gets is that the procedure, in present form, functions as a phenomenological information-pump picking up all "noticables," often in no apparent order. It is hard to get a sense of how important or salient different features are, or how they are related to each other. This version of the introspecting game should best be thought of as a method for collecting an exhaustive but unstructured inventory of stimulus aspects.

9 CONCLUSION

We began this paper by posing the following problem: How can a complete "outsider" determine whether an introspecting observer is successfully conveying information about internal events? Much of the paper was then devoted to establishing the formal properties of a procedure that made such a determination possible. In this, concluding section, I would like to state and briefly discuss three general conclusions that are suggested by this work:

(1) To understand instrospection as a form of communication, one needs to assume underlying structures of mutual knowledge, that model how different persons' internal variables are co-determined by stimulus presentation. In this paper, these structures were identified with joint probability distributions over internal states of two subjects.

(2) The exploration of mutual knowledge is not afflicted with any special methodological difficulties. In particular, there exist procedures that will instruct subjects, through numerical scores, how to identify and report any internal variable that correlates with at least one other variable, public or private.

(3) In the context of such procedures, introspection reappears as a cognitive skill, which we can define as the production of a comprehensible record of mental activity, in a personal vocabulary. As such, it may benefit from training, and from creative experimentation with different means of expression.

The assumption of mutual knowledge

Consider the following imaginary dialogue, between experimenter and a subject who is being instructed in direct magnitude estimation of loudness (Stevens, 1957):

Experimenter: "Please assign numbers to the tones that you will hear, in such a way that ratios of numbers equal ratios of their lodnesses."

Subject: "I am not sure exactly how to do this. Can you explain to me what exactly I am supposed to attend to in assigning these numbers?"

Experimenter: "I am afraid I can't do that: just use your own judgment."

Subject: "Can you at least show me some examples of good magnitude ratings?"

Experimenter: "I'm afraid I can't help you in that way either. Pick the numbers that reflect your judgment of magnitude -- not what you think other people have done. But, please be careful that the numbers you assign accurately reflect relative loudness."

If the subject now goes on and selects numbers in a very unusual manner, the experimenter will discount his responses as showing a lack of cooperation, attention, or understanding of the task. If all subjects produce idiosyncratic judgments, then the data will probably be declared incoherent, and not published. If on the other hand, the responses are related in

some striking way, as is the case with magnitude ratings, then that will be taken as evidence that subjects understood the instructions (on some level), and that the results may be integrated with the rest of psychophysical theory.

In current introspecting experiments, thus, the coherence of subjects' responses functions as a implicit criterion for separating meaningful from meaningless data. With few notable exceptions (such as the work on color names: Brown, 1976; Heider and Olivier, 1975; Lantz & Stefflre, 1964), however, psychologists have been reluctant to instruct subjects to aim for agreement -- to target their responses to the subjectively anticipated group mean. Let us speculate for a moment about why this is so.

The mutual consistency of subjects responses reflects, presumably, a measure of perceptual agreement, as well as a measure of agreement about the correct application of response categories to internal events. By explicitly instructing subjects to guess the mean response, for each stimulus, the experimenter runs the risk of having his subjects conceal perceptual differences through compensating adjustments in the interpretation of the response categories. Given this possibility, the experimenter prefers to leave his subjects without an explicit definition of correct performance, and hopes, in turn, that differences in responses will reflect primarily perceptual differences, rather than different understanding of the intentionally ambiguous verbal instructions.

The dilemma, as described here, evokes the problem of separating the effect of the so-called perceptual and decision factors in discrimination, which was solved successfully by signal-detection methodology. The research conducted in that framework showed that even a request as simple as

that of identifying the presence or absence of a weak stimulus, left the subject with a discretionary margin, i.e., the location of the boundary that separates phenomenal events into one or the other category. Surely, then, the problem of teasing out the influence of decision factors becomes more urgent when the instructions are as ambiguous as they are in most introspective tasks.

In this paper, we took the approach that whenever the introspecting responses of different subjects show some systematic inter-relationships (of which agreement is only the most simple kind), then one ought to postulate, as the source of this coherence, a structure of mutual knowledge that the subjects are tapping (but of which they are not necessarily aware). Although it may seem a bit unusual to speak of mutual knowledge of internal states and process, the mathematical formulation of this concept is straightforward, and is a direct consequence of any formal probabilistic model of the relation between stimulus and perceptual variables: If internal variables α,β of two subjects, A and B, are correlated with some stimulus parameter, $\underline{s}$, then they themselves are also correlated, and this correlation can be made a direct object of investigation, bypassing the linking physical parameters. The structure of mutual knowledge bears the same kind of relation to the observed patterns of intersubjective agreement, as, say, Thurstonian representations of sensory dispersion bear to patterns of stimulus discrimination. By approaching the mutual knowledge among two or more people as a direct object of investigation, we also avoid the restrictive assumption of isomorphy between people's internal states, that is often thought to be a prerequisite for meaningful introspection.

What are the limits of payoff-translation?

In 1961, Ward Edwards wrote an article entitled, "Costs and Payoffs are instructions," in which he explained why procedures that reward subjects according to some objective performance criterion have become the instruments of choice for modern experimental psychology. Edwards recognized three separate benefits that performance criteria provide. Their first (and perhaps least important) function is to reward, and so motivate subjects to an adequate level of attentiveness and effort. Second, they are a source of feedback that informs subjects how well they are doing. Third, they can resolve contradictions and ambiguities in verbal instructions, which arise, for example, when the verbal instructions do not specify a trade-off between potentially conflicting task requirements (such as speed and accuracy).

As the title of his article indicates, Edwards was primarily concerned with the third, communicative function of payoffs, as an unambiguous language for conveying instructions. But implicit in his article was a question that he did not address, namely, how far can one push this process of translation. Can all verbal instructions be replaced, in principle, by a scoring schedule? While it is hard to see how this question could be answered in full generality, the game presented in this paper shows that an important class of verbal instructions -- those that roughly correspond to requests for stimulus description -- does admit to such a translation, and that such requests for introspective information belong on the more secure side of this methodological fault line.

To gain perspective on our result, let us temporarily entertain a more abstract conception of the three-person game, as a procedure that

Insert Table 3 About Here

generalizes, and subsumes as special cases, previous methods of obtaining information about internal variables. The relevant comparisons are presented in Table 3, which is organized so as to show how other methodologies may be derived from the three-person game by substituting the experimenter for certain players, and/or by deleting some players altogether.

The first two lines in the Table draw a distinction between a phenomenological style of introspection, which encourages subjects to describe inner states in whatever terms seem appropriate, and the proper introspectionism of experimental psychology, which restricts subjects to a previously specified set of labels, scale ratings, and the like. Both variants are subjective, in that the subject has no external criterion by which to measure success.

In the signal detection procedure, (and, more generally, in any objective procedure), the experimenter takes over the role of Player A, since he defines the set of possible stimulus descriptions, as well as which particular description is correct on any trial. What the Table makes clear is that, starting with the signal-detection paradigm, we can build up the introspecting game by applying two procedural transformations. With the first transformation, we reverse, as it were, the roles of the subject and the experimenter, so that the subject is now entrusted with developing his own vocabulary for describing stimuli, and his success in this is measured by how well the experimenter can guess his responses, on the basis

Method	Examples	Definition of response set	Stimulus description	Informed (Posterior) Assessment	Uninformed (Prior) Assessment
1. Introspection (Free)	Phenomenology; projective tests.	---	Subject	---	---
2. Introspection (Regimented)	Direct magnitude scaling; similarity judgment.	Experimenter	Subject	---	---
3. Objective (Indirect)	Signal-detection paradigm.	Experimenter	Experimenter	Subject*	---
4. Objective (Direct)	None.	Subject*	Subject*	Experimenter	Experimenter
5. Three-person communication game.		Subject A*	Subject A*	Subject B*	Subject C*

*Responses guided by an explicit scoring system.

Table 3: A comparative summary of methods.

of public stimulus characteristics.[22] As far as I know, such a procedure (Method 4 in the Table) has never been used, which is perhaps unfortunate, since it places the subject in the active role of formulating a personal referential system, while at the same time avoids the strategic complications of the full-blown game.

The limitation of this type of procedure, of course, is that the scores are still conditioned on public stimulus characteristics, and will not reward identification of subjective aspects of experience that are uncorrelated with these characteristics. The three-person game removes this last limitation, by replacing the experimenter with a pair of subjects, one informed of the stimulus, and one informed of the stimulus set only, and establishing the differential accuracy of the informed subject over the uninformed one as the measure of information that the first subject should maximize.

Introspection as a skill

In this paper, we have presented a criterion for evaluating introspective reports, a criterion that is consistent with the statistical understanding of information as a reduction in uncertainty. Through their probability assessments, the two subjects, B and C, establish the two levels of uncertainty that are needed to measure uncertainty reduction; maximizing the gap between these two levels, as Subject A is enjoined to

---

[22] In such a procedure, Players B and C would be replaced with forecasting routines that assess probability distributions over A's responses on the basis of the relevant stimulus parameters, and A's response history.

do, is eqivalent to maximizing transmitted information. Although it would be premature to claim that the game which applies this criterion is the ideal instrument for collecting introspective information, its existence defines introspection as a skill, with objective standards of success and failure, and in this way clarifies the theoretical status of introspecting judgments that are obtained in more traditional ways.

Beyond this, however, our results suggest possibilities for a more flexible and creative use of introspection than is now considered appropriate. The main criticism of introspective observation, a criticism whose force is keenly felt whenever we encounter a descriptive system as unusual as the wine-drawings of Vandyke Price, is that there is no public mechanism for resolving the doubts of a skeptic -- a person who does not accept that the observations have any value. The communication game addresses this criticism directly, by instructing one subject to play the role of an in-house skeptic, whose score will decrease with every bit of genuine information that the other, introspecting subject reveals. Having thus installed the skeptical challenge right into the procedure, it is no longer necessary to limit subjects to descriptions of obvious and familiar stimulus aspects, nor, indeed, to the phenomenological repertoire that ordinary language provides. Rather, the game creates a setting in which new ways of describing mental activity can be developed and tested in a normal scientific manner.

REFERENCES

Abramson, N., Information Theory and Coding, McGraw-Hill, New York: 1963.

Aczél, J. and Pfanzagl, J., "Remarks on the measurement of subjective probability and information," Metrika, 1966, 11, 91-105.

Alpert, M., and Raiffa, H. "A Progress Report on the Training of Probability Assessors." In Judgment Under Uncertainty: Heuristics and Biases, D. Kahnemann, P. Slovic, and A. Tversky (Eds.). Cambridge, U.K.: Cambridge Univ. Press., 1982, 294-305.

Aumann, R. J., "Subjectivity and Correlation in Randomized Strategies," Journal of Mathematical Economics, 1974, 1, 67-69.

Aumann, R. J., "Agreeing to Disagree," The Annals of Statistics, 1976, 4, 1236-1239.

Aumann, R. J., "Correlated equlibrium as an expression of Bayesian rationality," Econometrica, 1987, 55, 1-18.

Brown, R. W., "Reference: In memorial tribute to Eric Lenneberg," Cognition, 1976, 4, 125-153.

Edwards, W., "Costs and Payoffs are Instructions," Psychological Review, 1961, 68, 275-284.

Garner, W. R., Uncertainty and Structure as Psychological Concepts, New York: Wiley, 1962.

Green, D.M., and Swets, J.A., Signal Detection Theory and Psychophysics, New York: Wiley, 1966.

Harasanyi, J. C., "Games with Incomplete Information Played by 'Bayesian' Players," Management Science, 1967-8, 14, 159-189, 320-334, 486-502.

Heider, E. R., and Olivier, D. C., "The Structure of the Color Space in Naming and Memory for Two Languages," Cognitive Psychology, 1975, 3, 337-354.

James, W., The Principles of Psychology, Cambridge, MA: Harvard University Press, 1981.

Kosslyn, S. M., Image and Mind, Cambridge, MA: Harvard Univ. Press, 1980.

Kullback, S., Information Theory and Statistics, Wiley, 1954.

Lantz, D., and Stefflre, V., "Language and Cognition Revisited," Journal of Abnormal and Social Psychology, 1964, 69, 472-481.

Lehrer, A., Wine and Conversation, Bloomington: Indiana Univ. Press, 1983.

Lewis, D. K. Convention, Cambridge, MA: Harvard Univ. Press, 1969.

Lichtenstein, S., Fischoff, B., and Phillips, L. D., "Calibration of Probabilities: The State of the Art to 1980," In Judgement Under Uncertainty: Heuristics and Biases, D. Kahnemann, P. Slovic, and A. Tversky (Eds.). Cambridge, U.K.: Cambridge Univ. Press., 1982-306-334.

Luce, R. D., and Raiffa, H., Games and Decisions, New York: Wiley, 1957.

Malt, B. C., and Smith, E. E., "Correlational Structure in Semantic Categories," cited in Categories and Concepts, E.E. Smith, and D. L. Medin, Cambridge, MA: Harvard University Press, 1981.

Mertens, J. F., and Zamir, S., "Formalization of Bayesian Analysis for Games with Incomplete Information," International Journal of Game Theory, 1985,14, 1-29.

Milgrom, P., "An Axiomatic Characterization of Common Knowledge," Econometrica, 1981, 49, 219-222.

Nash, J., "Non-Cooperative Games," Annals of Mathematics, 1951, 54, 286-295.

Nisbett, R.E., and Wilson, T.D., "Telling more than we know: Verbal reports on mental processes," Psychological Review, 1977, 79, 231-279.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P., "Basic Objects in Natural Categories," Cognitive Psychology, 1976, 8, 382-439.

Savage, L. J., The Foundations of Statistics, New York: Wiley, 1954.

Savage, L. J., "Elicitation of Personal Probabilities and Expectations," Journal of the American Statistical Association, 1971, 66, 783-801.

Shannon, C. E., Weaver, W., The Mathematical Theory of Communication, Urbana: Univ. of Illinois Press, 1949.

Shepard, R. N., "Representation of Structure in Similarity Data: Problems and Prospects," Psychometrika, 1974, 39, 373-421.

Skinner, B.F., "The operational analysis of psychological terms," Psychological Review, 1945, 52.

Stevens, S.S., "On the psychophysical law," Psychological Review, 1957, 64, 153-181.

Swets, J. A., Tanner, W. P., Jr., and Birdsall, T., "Decision Processes in Perception," Psychological Review, 1961, 68, 301-340.

Toda, M., "Measurement of Subjective Probability Distribution," Report No. 3, Division of Mathematical Psychology, Institute for Research, State College, Pennsylvania, 1963.

Tversky, A., and Gati, I., "Similarity separability, and the triangle inequality," Psychological Review, 1982, 89, 123-154.

Tversky, A., and Kahneman, D., "Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgement," Psychological Review, 1983, 90, 293-315.

Vandyke Price, P., The Taste of Wine, New York: Random House, 1975.

Appendix I

Proposition 5 states that if all three subjects play optimally, then the score in the sequential game does not depend on the number of stages allowed, nor on the order in which information is revealed by Player A. We now sketch the proof of this for a two-stage game; the inductive generalization to more complex games is straightforward.

Let $\underline{a}_i$ and $\underline{A}_i$ (i=1,2) designate the descriptions and vocabularies selected by A, in Stages 1 and 2. The strategies for A, B, C are given below:

	Stage 1	Stage 2
Player A	$r(\underline{a}_1,\underline{A}_1,\underline{a}_2,\underline{A}_2,\|\alpha)$	
Player B	$b(\underline{a}_1\|\underline{A}_1,\beta)$	$b(\underline{a}_2\|\underline{A}_2,\underline{a}_1,\underline{A}_1,\beta)$
Player C	$c(\underline{a}_1\|\underline{A}_1)$	$c(\underline{a}_2\|\underline{A}_2,\underline{a}_1,\underline{A}_1)$

Notice that Stage-2 strategies of B and C are conditioned on all the information that has been revealed up to that point. As before, we extend A's strategy, r, to be a distribution over β,

$$r(\underline{a}_1,\underline{A}_1,\underline{a}_2,\underline{A}_2,\alpha,\beta) \equiv r(\underline{a}_1,\underline{A}_1,\underline{a}_2,\underline{A}_2|\alpha)p(\alpha,\beta).$$

The equation for expected score is a weighted sum of scores in the two stages:

$$EV(r,b,c) = \sum r(\underline{a}_1,\underline{A}_1,\underline{a}_2,\underline{A}_2,\alpha,\beta)\left[\log\frac{b(\underline{a}_1|\underline{A}_1,\beta)}{c(\underline{a}_1|\underline{A}_1)} + \log\frac{b(\underline{a}_2|\underline{A}_2,\underline{a}_1,\underline{A}_1,\beta)}{c(\underline{a}_2|\underline{A}_2,\underline{a}_1,\underline{A}_1)}\right].$$

The trick, again, is to expand the expression inside the logarithms into a product, composed now of eight (rather than five) ratios:

$$p(\beta|\alpha)/p(\beta), \tag{R1}$$

$$p(\beta)/r(\beta|\underline{A}_1), \tag{R2}$$

$$r(\beta|\underline{a}_2,\underline{A}_2,\underline{a}_1,\underline{A}_1)/p(\beta|\alpha), \tag{R3}$$

$$b(\underline{a}_1|\underline{A}_1,\beta)/r(\underline{a}_1|\underline{A}_1,\beta), \tag{R4}$$

$$r(\underline{a}_1|\underline{A}_1)/c(\underline{a}_1|\underline{A}_1), \tag{R5}$$

$$b(\underline{a}_2|\underline{A}_2,\underline{a}_1,\underline{A}_1,\beta)/r(\underline{a}_2|\underline{A}_2,\underline{a}_1,\underline{A}_1,\beta) \tag{R6}$$

$$r(\underline{a}_2|\underline{A}_2,\underline{a}_1,\underline{A}_1)/c(\underline{a}_2|\underline{A}_2,\underline{a}_1,\underline{A}_1), \tag{R7}$$

$$r(\beta|\underline{a}_1,\underline{A}_1)/r(\beta|\underline{A}_2,\underline{a}_1,\underline{A}_1). \tag{R8}$$

Careful application of Bayes' rule shows that the distributions r and p drop out when the terms (R1)-(R8) are multiplied together. By inserting this long product into the equation for expected score, we obtain a definition of expected score as a sum and difference of eight expected divergences, very much like the ones derived in Section 4. The divergences corresponding to R4, R5, R6, and R7, define B's and C's optimal assessments, which are indeed their true subjective assessments. The divergences corresponding to R2, R3, and R8 give three conditions for A's optimal strategy:

$$r^\circ(\beta|\underline{A}_1) = p(\beta), \qquad \text{(from R2)}$$

$$r^\circ(\beta|\underline{a}_2,\underline{A}_2,\underline{a}_1,\underline{A}_1) = p(\beta,\alpha) \qquad \text{(from R3)}$$

$$r^{\circ}(\beta|\underline{A}_2,\underline{a}_1,\underline{A}_1) = r^{\circ}(\beta|\underline{a}_1,\underline{A}_1) \qquad \text{(from R8)}$$

The first condition is the same as in the single-stage game; the third condition is entirely analogous: it states that the second vocabulary should not provide any information about β, beyond what has already been revealed by $\underline{a}_1$, and $\underline{A}_1$. The middle condition, taken from R3, states that the two pairs of vocabularies and descriptions must reveal all information in α that is relevant about β, but, significantly, it doesn't tell A how to divide this information between the first and second stage of the game. If the three players play optimally, then all of the divergences will equal zero, except the one corresponding to R1, which yields the same expected score, $I(\alpha:\beta)$, as the single-stage game.